

E

Earth System Environmental Literacy

Sustainability Education: Challenges of Integrating Virtual Versus Real Nature in Science Education

MARGARET LOWMAN

Nature Research Center, NC Museum of Natural Sciences and Professor, North Carolina State University, Raleigh, NC, USA
TREE Foundation, Sarasota, FL, USA

Article Outline

Glossary
Definition of Sustainability Education
Introduction
Short History of Sustainability Education and Related Education Initiatives
Sustainability Education in the Twenty-First Century
Future Directions
Bibliography

Glossary

Ecosystem services Benefits provided by natural systems that do not necessarily require significant financial maintenance, including water purification, clean air, pollination, soil decomposition, biodiversity, and many others. In accounting, balance sheets often overlook these free services provided by Mother Nature.

Finite resources In sustainability education, this involves those natural resources on planet Earth that require careful allocation into the future, such as fresh water, oil, and many others.

NEON Acronym for a long-term environmental monitoring program funded by Congress and mandated to include both research and education components, entitled “National Ecological Observatory Network.”

Pedagogy The art, science, or profession of teaching and indirectly, the process of implementing assessments or outcomes as part of educational research in order to improve teaching.

STEM Acronym describing the major science subjects for education advancement, which include “science, technology, engineering, and mathematics.”

Sustainability When a system functions indefinitely with regard to present and future needs, such as when human and natural systems operate indefinitely through careful resource allocation.

Virtual versus real In sustainability education, there is an underlying controversy between the importance of tools of teaching: newer techniques involve virtual tools (computers, handheld devices, simulated games, virtual field trips, and videos), whereas more conventional tools include real nature such as field trips, taxonomic collections, and hands-on natural science activities. Ideally, both virtual and real approaches contribute to effective learning.

Definition of Sustainability Education

- The best time to plant a tree is twenty years ago. The next best time is now. (African proverb)

If every citizen could read the above quote and understand its underlying ecological concepts, economic challenges, social services, and spiritual heritage, then it is likely that sustainability education would be achieved. The notion of a tree and its ecosystem services illustrate sustainability in the simplest yet most robust sense. To plant and grow a tree, economists struggle with volatile currencies; ecologists juggle development and conservation; religious leaders advocate stewardship; and social scientists examine equity in a world of declining resources. Sustainability education requires an integrated approach between ecology, risk analyses, economics, social sciences, biological sciences, political sciences, languages, biotechnology, physical sciences, health sciences, and religion.

All these practitioners (and many others) contribute to sustainability education, an emerging discipline that requires an interdisciplinary synthesis of knowledge, translated into practice, to insure the future of life on Earth.

Introduction

Science education, also termed STEM (science, technology, engineering, and mathematics) is undergoing a revolution in America as well as overseas [1]. The clear links between our environment, human health, and economics have triggered a serious effort to improve science literacy into the twenty-first century. As global environmental setbacks including climate change, famine, and biodiversity extinction loom, Earth's living systems are fast approaching a tipping point beyond which life on earth may cease to exist as it is known now [2]. Yet never before have humans had such high levels of technology to achieve solutions to global challenges. How to integrate these new technologies with science education is one of the major priorities facing both local and national governments. Neighborhoods and communities face the challenges of air quality and fresh water; entire nations prepare to adapt to rising sea levels and volatile food prices due to droughts; and around the globe, nations struggle with increasing levels of carbon dioxide in our atmosphere and oceans [3]. All of these issues directly link to the quality of sustainability education, which translates into training the next generation to understand the links between a healthy environment, human health, a sound economy, and shared use of finite natural resources.

It is daunting to link the future health of planet Earth with the importance of learning about how our Earth operates. But science or sustainability education in this sense has become a direct metric for our future survival. In America, professional groups including National Academy of Sciences, National Science Foundation, American Association for Advancement of Science, National Association of Environmental Educators, Boy and Girl Scouts of America, American Association of Museums, almost all scientific societies, and even grassroots organization such as Children and Nature Network, local Sierra Club chapters, and religious networks have embraced the challenge of raising

science literacy. Furthermore, science literacy is adopted here to umbrella many different aspects of STEM literacy. Diverse components of science literacy included under this umbrella range from marine literacy, environmental literacy, physics literacy, mathematics literacy, and to ecological literacy. Sustainability goes one step further from the collective science literacies in that it requires the integration of both social and physical/biological sciences. But one common denominator emerges for all of these efforts relating to science or sustainability literacy – the notion of sustainability education embraces our knowledge of how to continue to live on planet Earth without exhausting our resources.

Ultimately, sustainability education may be best achieved by a blend of hands-on (i.e., real) and virtual science education experiences. Building bridges between virtual and real environments, among scientists and citizens, and spanning ecology and economics were some of the central issues in recent summits and conferences of educators and other professionals in Washington, DC during 2009–2010. This chapter reviews some of the emerging informal science education priorities for integrating tools both virtual and real, ultimately to create effective sustainability education.

Short History of Sustainability Education and Related Education Initiatives

Over the past decade, myriad conferences, conversations, workshops, and white papers have been produced on the topic of science education, and on raising America's STEM literacy as part of its global economic competitiveness [4–13]. It is well recognized that the conventional methods of large lectures, textbooks, and short laboratory sessions are not able to cover the overwhelming body of scientific knowledge, and even worse, this format does not inspire today's students to pursue careers in science or even to achieve a minimal level of scientific literacy. Educators and scientists alike recognize that innovative, new pedagogies need implementation immediately, if America can regain her leadership in technology and innovation, especially in fields that relate to sustainability such as energy, conservation, biodiversity, ecosystem services, and environmental justice. This commitment

has gained center stage in editorials, publications, and conversations from the highest federal levels to the grassroots of small school PTAs; yet specific mechanisms for achieving a robust level of sustainability education have not yet been successfully documented. This remains a priority for both the current practitioners of STEM education, for educators, and for the stewards of natural resources, who recognize the need for interdisciplinary approaches in education to ensure sustainability for all.

Ecological Literacy

- There is one important difference between ecology and many other fascinating sciences and games: unsolved problems of chess, astronomy, or mathematics will not change if we ignore them. Our activity or lack of activity can alter the state of ecology. (Lawrence B. Slobodkin 2003)

Ecological or environmental literacy has been a cornerstone for both formal and informal science education [14]. In particular, the big challenge for ecology education is not a lack of information, but rather the need to provide the relevant context [15, 16] that will motivate the next generation of environmental scientists to collect, access, and interpret relevant information needed for ecological stewardship. A conceptual understanding of “healthy” ecosystems and related ecosystem services, ranging from food and energy to clean air and water, will be required – not only to serve as the foundation for sound economics, but also to sustain human health. Nature shapes, and is shaped, by communities where people live, and in this sense ecology education overlaps directly with sustainability education. When students bond with their natural surroundings, and then use virtual simulations to understand large-scale ecological processes and drivers, ecology education is achievable [17, 18].

During October 2010, a diverse group of organizations hosted a sustainability education summit, entitled Environmental Literacy for a Sustainable World (www.esa.org/eesummit). The mission of this summit was *not* to create a white paper to mail out to participants 6 months later, as is often the aftermath of conferences, but to create a decadal action plan embraced and activated by the participants to advance environmental literacy. One of the first notions of the summit

was to agree that all the science literacies are closely related, and that the achievement of a sustainable world will only come about by raising the standards of ecological, environmental, and science literacy – or whatever the label. An open letter to government leaders about the importance of sustainability education was one of the first outcomes (for, e.g., 17 January 2011, www.heraldtribune.com). Additional action items include: (1) Produce a common set of 5–15 key concepts and principles that link human and natural systems for use by teachers and policy makers; (2) Engage communities to meet the challenges of an uncertain future; (3) Create a network of green-design built environments; create an “APP” (application for handheld mobile devices) that links personal behavior to the environment as an index of health; (4) Construct a “go-to” site for environmental materials, aka a technology infrastructure to inventory, synthesize and share best practices and programs; and others. These action items all have different timelines, with the ultimate goal of engaging diverse stakeholders in different elements of sustainability education by 2020.

Conservation Literacy

Like ecology, conservation biology represents another closely related discipline to sustainability, and hence the concepts and applications of an educational strategy will overlap. In the field of conservation biology, a lively debate about the creation of a conservation ethic started long ago with Aldo Leopold’s seminal book, *The Sand County Almanac* [14]. Leopold articulated a value system with regard to land, “That land is a community is the basic concept of ecology, but that land is to be loved and respected is an extension of ethics. That land yields a cultural harvest is a fact long known, but latterly often forgotten.”

The Society for Conservation Biology defined its goals and objectives as, “the education, at all levels, preparatory and continuing, of the public, of biologists, and of managers in the principles of conservation biology” [19]. Further, the Society’s education committee articulated principles that seek to maintain Earth’s biological diversity, ecological integrity, and ecological health. But the practitioners of conservation biology admit that their approach is very passive,

“Among conservation biologists the prevailing view of conservation education appears to rest on

a deep-seated belief that the basic function of conservation education should be to disseminate knowledge that scientists generate, essentially to transport information to the public and key groups in the expectation that it will eventually precipitate more appropriate conservation-related behaviors” [20]. This is not unusual, and most scientists have adopted a “speak and they will listen” approach to STEM education. But in a world of declining and finite resources, a tipping point is fast being approached in the education circles, whereby the methodologies need to drastically change in order to inspire a science-literate community as stewards of global health. Sustainability education – if launched with appropriate integration of virtual and real natural systems [21], facts and “oh-wow” responses, local and global perspectives, and interdisciplinary approaches [22, 23] – is critical to succeed.

Biological Literacy

As a biologist by training, I have tracked the notion of biological literacy as a component of science education throughout my career [1]. The world has witnessed an amazing explosion of knowledge in biology – ranging from genomics and nanotechnology to global climate change and how sea-level rise will impact Earth’s resources.

The relevance of biology education may be the most focused of all the “science literacies” to date due to extensive support from NSF, AAAS, Association of American Medical Colleges (AAMC), and Howard Hughes Medical Institute (HHMI). In its most recent “conversation” that included over 500 biological sciences faculty, administrators, students, and other practitioners, AAAS published a white paper called “Vision and Change” (www.visionandchange.com) [1]. They summarized four academic concepts to implement change in biology education: (1) Integrate core concepts and competencies; (2) Focus on student-centered learning; (3) Promote a campus-wide commitment to change; and (4) Engage the biology community in the implementation of change. The outcome of the “conversation” was to encourage all biology educators to develop a coordinated and sustainable plan for implementation of

principles that will improve the national standards of biology education.

Specific working groups to advance biology education (also termed “new biology”) convened around different priorities needed for biology education: managing large, complex data sets; interdisciplinary approaches; making connections among different concepts and/or questions; and seek new inquiry-based learning activities rather than the conventional memorization and lecture-based approaches [1]. Advancing education in biological sciences directly impacts our nation’s ability to respond to recent challenges, ranging from climate change to safe drinking water [10–12]. The core literacy concepts were delineated to include evolution, structure and function, information flow, and pathways and transformations of energy/matter, systems [1]. Core competencies were defined as the abilities to apply the process of science, use quantitative reasoning; use modeling/simulation; tap into the interdisciplinary nature of science; communicate/collaborate with other disciplines; and understand the relationship between science and society [1]. While these criteria were developed for biology education, many of the same elements would be relevant for sustainability education which concurrently embraces biology as well as other disciplines.

Sustainability Education in the Twenty-First Century

Achieving science literacy is vital in the twenty-first century. Sustainability education can be construed as that interdisciplinary platform whereby diverse members of global communities seek to understand the role of humans in their physical and social environments. As global environmental challenges loom at an unprecedented magnitude, “tipping points” that threaten to irreversibly damage Earth’s living systems are fast approaching [16]. Yet, never before have humans had such a wealth of technology to achieve solutions at their disposal, allowing collaboration from virtually anywhere in the world, drawing ideas from multiple disciplines to process and analyze countless data points, and teaching the next generation to view our world in exciting, novel ways that inspire environmental stewardship.

In our teaching, environmental scientists must seek to balance cellular versus organismal biology, virtual models versus real-time data, and science blended with policy. Future stewards will need skills in assessing, predicting, managing, and communicating the dramatic ecological and societal risks and changes confronting us [1]. However, a major stumbling block in training the next generation of practitioners is the challenge of effectively integrating technology with in situ fieldwork. While most senior ecologists were inspired by their training in the field, younger scientists are more familiar with virtual ecosystems through gaming, social networking, and computer models, sometimes leading to so-called nature-deficit disorder. How can environmental practitioners blend “hands-on” fieldwork with “cutting-edge” technology? This conundrum is the subject of ongoing debate. Those born after 1980 spend more time indoors with electronic devices than going outdoors to experience nature first hand [17, 18].

On a positive note, new programs are emerging that successfully integrate virtual and real environments. The National Ecological Observatory Network (NEON) will conduct continental-scale environmental monitoring in situ, and their large databases will be accessible to students, citizen scientists, and policy makers ([20], see also www.neoninc.org). At the North Carolina Museum of Natural Sciences in North Carolina, the new Nature Research Center, with its mission “to engage the public in understanding the scientific research that affects their daily lives,” will house state-of-the-art research laboratories, accessible to the public, a three-story Daily Planet immersion theater, broadcasting field science from remote sites via video-streaming, and dedicated virtual and real meet-the-scientist activities ([21], see also www.naturesearch.org). These two examples illustrate the changing landscape for ecology education, and how technology can, and must, advance environmental literacy.

NEON

The National Ecological Observatory Network will take field measurements by scientists, citizens, and students collectively to a new level [24]. This continent-scale research platform will conduct long-term monitoring

in America’s major ecosystems, and create an important education outreach component as a parallel stream of the research program. Affectionately called the “ecologists’ Hubble Telescope,” this is the largest federally funded research project undertaken by the ecological community, and aspires to contribute data sets that will lead to best practices for solving environmental challenges in the future. Critical to NEON’s success is the sustainability education and outreach component that have the potential to change how policy makers as well as citizens learn about sustainable solutions.

Citizen science has burgeoned in the past decade, in part due to Internet and handheld technological devices, but also because of a growing awareness that sustainability education is a community responsibility. Almost 20 years ago, the Jason Expedition (www.jason.org) pioneered distance learning for middle school youth, who not only visited marine and rain forest canopy environments, but also collected ecological data to share with other schools (e.g., [25]). Today, myriad programs including Project Budburst (www.budburst.org) and Project Nestwatch (www.nestwatch.org) rely on citizens’ collecting data, which are collated at a central source for scientific use.

Planning for NEON took over a decade with hundreds of ecologists providing input. Projected to become fully operational by 2016, NEON’s 20 field sites will be linked by high-capacity cyber infrastructure with the capacity to make forecasts about environmental changes. Important for sustainability education, these issues include six of the National Academies’ grand challenges: biodiversity, climate change, ecohydrology, biogeochemistry, infectious disease, invasive species, and land use [26]. And even more critical to sustainability education is NEON’s capacity to involve not just scientists – but also the public, students, teachers, and policy makers – to use the NEON interface for downloading data and learning by doing. NEON also pledges to prioritize science communication to diverse audiences [24] and foster the links between human health and healthy ecosystems. Will NEON’s education vision be achieved, or will it become a pile of dusty papers on some university desks? Only time will tell, but NEON has the potential to advance sustainability education across an entire continent.

NRC

The Nature Research Center (NRC) at the North Carolina Museum of Natural Sciences is a new, 80,000-square foot wing of the existing 132-year-old Museum. While conventional museums focus on “what we know” about science, the new wing will engage the public in “how we know what we know” and in particular inspire the public to understand scientific research that affects their daily lives. The NRC goals are to conduct/communicate cutting-edge science, create exemplary role models to inspire diversity in the next generation of scientists, and integrate real and virtual science to engage youth about scientific discoveries. Four new labs will house scientists, educators, technology experts, students, and volunteers – all working together to understand how science works and communicate it to a diverse audience. In addition, the notion of scientists and educators and technology geeks all working together on the floor is unique to the Museum world, and seeks to assimilate “citizen science” into all components of research and education at the NRC. Many partners will insure a robust content for science communication – joint positions with the

UNC system, partnerships with NOAA, NASA, National Geographic, and North Carolina government agencies, will all contribute to the creation of innovative science communication. The NRC has four unique learning platforms, facilitating the assessment of how different audiences learn best: (1) The Daily Planet, a high technology, immersion theater linked through the Ethernet highway to all classrooms in North Carolina and ultimately, around the world; (2) Citizen science labs, where participation in data collection by volunteers will contribute to answering important questions about our environment; scientific laboratories with glass walls allowing visitors to be involved in the research; (3) Investigate labs, where related research will be undertaken that involves school groups or other visitors; and (4) Hands-on exhibits where visitors are immersed in *how science works* (Fig. 1).

By “reinventing” the education pathways of conventional museums, the NRC hopes to inspire diverse students to become excited about science, and to inspire scientists to engage in successful education outreach, which links directly to creating a higher level of sustainability education.



Earth System Environmental Literacy. Figure 1

North Carolina's new Nature Research Center integrates cutting-edge research, education and technology for diverse audiences to foster environmental literacy

Future Directions

The future is ours by design, but there is a need to seek interdisciplinary solutions and then communicate them with one voice. When baby boomers think back to their childhood, they can probably recall a favorite tree house, scout camping trip, or neighborhood picnics. In today's world, children are not allowed unsupervised time outdoors, due in part to parents' distrust of issues ranging from strangers to global security. But knowledge of real nature is essential to raise levels of environmental literacy, and in turn achieve a higher degree of sustainability education. Harking back to the African proverb in the beginning of this entry, many children do not have a clear concept of what is required to sustain a tree. But knowledge of nature is their best weapon if young people are ultimately to make good decisions about personal health, climate change, and land-use management.

Despite the gloom and doom notions surrounding environmental literacy and our human aspiration to attain a sustainable lifestyle, there are exemplary actions providing amazing examples of best practices. What are some of the diverse projects exemplifying

sustainable education? Here is a selection of sustainability education activities, perhaps useful to implement in your own community!

1. *Strategies for Ecology Education, Diversity and Sustainability (SEEDS). Diverse People for a Diverse Science.* SEEDS is an education program of the Ecological Society of America (Fig. 2). Its mission is to diversify and advance the ecology profession through opportunities that stimulate and nurture the interest of underrepresented students to not only participate in ecology, but to lead. Focused mainly at the undergraduate level, with extension services for communities, high schools, graduate students, and international collaborations, the SEEDS program promotes an ecology profession with wide representation to ensure environmental understanding and a sustainable future for all. (www.esa.org/seeds)
2. *Urban Farming.* In partnership with a neighboring community garden, Mill Creek Farm (MCF) transformed a vacant West Philadelphia lot into a thriving urban farm and intergenerational educational center (Fig. 3). Now starting on their sixth



Earth System Environmental Literacy. Figure 2

At the Ecological Society of America's annual meetings, SEEDS students are mentored to insure a diverse community of ecologists



Earth System Environmental Literacy. Figure 3

Mill Creek Farm, near Philadelphia Pennsylvania, exemplifies the growing popularity of sustainable urban farming

field season, MCF engages and galvanizes the community by providing affordable and sustainable food to the surrounding neighborhoods, employing local high-school students, offering food justice education and much more. For more information (including a documentary and radio stories) visit: <http://www.millcreekurbanfarm.org/>

3. *Sustainability Education for Incarcerated Youth*. For 19 years, high school science teacher DC Randle taught biology and environmental sciences to middle and high school youth in the Minnesota correction system. Defined as emotionally, behaviorally disordered students, they thrived under a curriculum that allowed them to rediscover nature. Using the natural resources around their facility, their science teacher provided a strong backbone for sustainability education to these at-risk youth.
4. *Local-to-Global Connections*. As part of a fundraiser for a local canopy walkway, Meg Lowman started leading field trips to the Amazon for underserved youth in Florida. Called the Florida-Peru Connection, the local community funded scholarships for local Boys and Girls Club and Girls Inc. students. In return, young people have visited America from Amazonian villages along the Rio Napo in Peru. The link between local and global is an important concept for sustainability, and experiencing other ecosystems and culture first hand had a significant impact on the perspectives of both the Amazon and the American communities who were part of the exchange. As a result of shared global views, both Florida and the Amazon share a canopy walkway for education, where children (and others) conduct research on the biodiversity of forests (Fig. 4). More information is listed at www.treefoundation.org
5. *Student Outreach in Science (SOS)*. Environmental studies students at New College of Florida developed sustainability outreach activities appropriate for elementary and middle school classes. They developed a skillset in science communication, and then took their lessons to underserved schools in the districts, providing a student-teaching-student model that was well received by both the teachers and the students. Units ranged from climate change adaptation to growing local vegetables to invasive species to clean energy. Over 60 students at New College received the Sarasota County Conservation Education award based on their SOS program (www.canopymeg.com)



Earth System Environmental Literacy. Figure 4

At North America's first public canopy walkway in Florida, third graders conduct original research on forest canopies

6. *Master of Arts in Sustainable Communities, Northern Arizona University* (<http://home.nau.edu/sus/>). Students are admitted from all kinds of backgrounds – from political science and sociology to science and engineering to humanities and business. Students build their own programs of coursework from the university's very broad array of courses relevant to sustainability (of economies and of societies as well as of natural resources and environment). More important, they each develop independent, community-based research/thesis projects. Many come to this program as working professionals while some are more conventional students coming straight from a bachelor's degree. Graduates from this program are indeed transforming many communities around the intermountain west. Another unique program at NAU is the Professional Science Master's in Climate Science and Solutions. <http://climatesciencesolutions.nau.edu/>
7. *Measuring the Sustainability of Food to Table*. Professor David Cleveland at University of California, Santa Barbara engaged his environmental studies students in measuring the energy budget of regional agriculture, from farm to table. This agrifood systems analysis has the potential to change small-scale and local agriculture, and educate students in sustainable practices at the same time. It also earned David the award of Campus Sustainability Champion [27]. <http://es.ucsb.edu/profile/cleveland>
8. *Young Voices on Climate Change*. Author Lynn Cherry, finding resistance to teaching climate science, is producing inspiring, empowering short films to break down those barriers. The *Young Voices on Climate Change* documents inspiring success stories of kids shrinking their carbon footprint and fighting climate change. When climate change science is preceded with the showing of these films, youth and adults alike are more receptive to embracing the seriousness of environmental change rather than denying. Young people listen to others their own age and, in these films their peers model the skills, motivation, and confidence necessary to reduce CO₂ emissions. (www.YoungVoicesonClimateChange.com)
9. *Sustainability between Connecticut and Jamaica*. Eastern Connecticut State University recently had approved a new major in Sustainable Energy Studies – one interesting program associated with

students in this new major has been the service-learning project taking students to Jamaica to develop a small wind power system for the school in Lucea, Jamaica. The university has a long-standing relationship with Jamaican public schools because of our program to get their normal school teachers to complete their bachelor's degree in USA. This type of global experience in sustainability education is part of Eastern's two Strategic Plan Initiatives – Liberal Arts Work and Global Citizenship. Check out: http://www.diversityweb.org/DiversityDemocracy/vol14no2/loxsom.cfm?utm_source=pubs&utm_medium=blast&utm_campaign=divdemvol14no2

10. *The Perennials Collaborative*. The emerging generation of sustainability leaders brings new ideas for sustainability initiatives. Samir Doshi at University of Vermont and colleagues have undertaken a collective, action-oriented community project to inspire “Perennials,” those visionary leaders who (like perennial plants) have roots in communities and become forces of vision and change. Doshi's team is making a documentary, web site, and hosting community workshops throughout the country, especially in at-risk areas such as Appalachia, to inspire more actions for sustainability. (www.perennialscollaborative.com).

America is at a crossroads for education in science, technology, engineering, and mathematics (STEM). Currently, a company can hire nine Mexicans or eight Indian engineers at the same cost as one American; 93% of middle school students learn physical sciences from a noncertified teacher; and Americans spend more on potato chips than the entire federal R&D budget for energy [6]. It is no wonder that America currently ranks 48th in terms of global STEM education. Recent reports by the National Academies of Science, Engineering and other focus groups have assessed America's current STEM education, and listed actions that federal policy makers could take to enhance STEM and allow America to increase her global competitiveness [5, 6]. Such actions include increasing America's economy by vastly improving K-12 STEM education; doubling federal investments in STEM research to fuel our economy, provide security, and foster sustainability; and encourage citizen science to raise STEM

literacy. Will these actions be implemented, or will the Academies' report be overlooked? Only time will tell, but STEM and ultimately sustainability education collectively will invariably determine the quality of life for the next generation.

Bibliography

Primary Literature

1. Brewer CA, Smith D (2010) Vision and change in undergraduate education: a call to action. American Association for the Advancement of Sciences, Washington, DC, p 79
2. Pooley E (2010) The climate war. Harper Collins, New York
3. Braasch G (2007) Earth under fire. University of California Press, Berkeley, CA, USA, 286 pp
4. Council on Undergraduate Research (CUR) (2007) Developing and sustaining a research-supportive curriculum: a compendium of successful practices. CUR, Washington, DC
5. National Academy of Sciences (NAS) (2008) Rising above the gathering storm. National Academy of Sciences Press, Washington DC, USA
6. National Academy of Sciences (NAS) (2010) Rising above the gathering storm, revisited: rapidly approaching category 5. National Academies Press, Washington, DC
7. National Research Council (NRC) (1996) National science education standards. National Academies Press, Washington, DC
8. National Research Council (NRC) (1997) Science teaching reconsidered: a handbook. National Academies Press, Washington, DC
9. National Research Council (NRC) (1999) Transforming undergraduate education in science, mathematics, engineering, and technology. National Academies Press, Washington, DC
10. National Research Council (NRC) (2010a) Research at the intersection of the physical and life sciences. National Academies Press, Washington, DC
11. National Research Council (NRC) (2010b) Synthetic biology – building a nation's inspiration: interdisciplinary research team summaries. National Academies Press, Washington, DC
12. National Research Council (NRC) (2010c) Expanding underrepresented minority participation: America's science and technology talent at the crossroads. Committee on Underrepresented Groups and the Expansion of the Science and Engineering Workforce Pipeline. National Academies Press, Washington, DC
13. National Science Foundation (NSF) (1996) Shaping the future: new expectations for undergraduate education in science, mathematics, engineering, and technology. NSF, Washington, DC
14. Leopold A (1949) A sand county almanac. Oxford University Press, New York, NY, USA
15. National Research Council (2000) How people learn: brain, mind, experience, and school. National Academies Press, Washington, DC

16. National Science Foundation (2009) Advisory Committee for Environmental Education and Research. NSF, Washington, DC
17. Louv R (2006) Last child in the woods – saving our children from nature-deficit disorder. Algonquin Press, Chapel Hill
18. Lowman MD, Burgess E, Burgess J (2006) It's a jungle up there – more tales from the treetops. Yale University Press, New Haven
19. Trombulak SC, Omland KS, Robinson JA, Lusk JJ, Fleischner TL, Brown G, Domroese M (2004) Principles of conservation biology: recommended guidelines for conservation literacy from the education committee of the Society for Conservation Biology. *Conserv Biol* 18:1180–1190
20. Bride I (2006) The conundrum of conservation education and the conservation mission. *Conserv Biol* 20:1337–1339
21. Lowman M, Mourad T (2010) Bridging the divide between virtual and real nature. *Front Ecol Environ* 8(7):339
22. Balmford A, Cowling RM (2006) Fusion or failure? The future of conservation biology. *Conserv Biol* 20:692–695
23. Brewer C (2006) Translating data into meaning: education in conservation biology. *Conserv Biol* 20:689–691
24. Lowman M, D'Avanzo C, Brewer C (2009) A national ecological network for research and education. *Science* 323:1172–1173
25. Burgess E, Burgess J, Lowman MD (2003) Lowman and thousands of Jason X school students. 2003. Observations of a Coleopteran herbivore on a bromeliad in the Peruvian Amazon. *J Bromel Soc* 53:221–224
26. National Research Council (2001) Grand challenges in environmental sciences. National Academies Press, Washington, DC
27. Cleveland DA, Radko CN, Muller NM, Watson TD, Rekstein NJ, Van H, Wright M, Hollingshead SE (2011) Effect of localizing fruit and vegetable consumption on greenhouse gas emissions and nutrition, Santa Barbara County. *Environ Sci Technol* 45:4555–4562

Earth System Model, Modeling the Land Component of

GUO-YUE NIU¹, XUBIN ZENG²

¹Biosphere 2, University of Arizona, Oracle, AZ, USA

²The Department of Atmospheric Sciences, University of Arizona, Tucson, AZ, USA

Article Outline

Glossary

Definition of the Subject and Its Importance

Introduction

Major Processes Represented in an LSM

Future Directions

Acknowledgments

Bibliography

Glossary

Aerodynamic resistance Or drag or aerodynamic drag. The component of force exerted by the air on a liquid or solid object that is parallel and opposite to the direction of flow relative to the object.

Evapotranspiration A term used to describe the sum of evaporation and plants' transpiration from the Earth's land surface to the atmosphere.

LAI The ratio of total upper leaf surface area of vegetation divided by the surface area of the land on which the vegetation grows. LAI is a dimensionless value, typically ranging from 0 for bare ground to 6 for a dense forest.

Land surface processes The various biogeophysical and biogeochemical processes occurring within and over various land surface components and interacting with the atmospheric processes.

Parameterization A method or process to approximate the effect on the model resolved processes that would be generated by unresolved subgrid-scale processes using formulations containing empirical parameters.

Stomatal resistance The opposition to transport of quantities such as water vapor and carbon dioxide (CO₂) to or from the stomata of plant leaves.

Subgrid heterogeneity The spatial variation in land surface properties, for example, soil wetness, snow, vegetation, and terrain within an ESM grid cell, which is on the order of ~100 km.

Tile method A method dealing with subgrid heterogeneity. It gathers similar homogeneous land patches at different geographic locations within an ESM grid-cell into a number of "tiles" or "mosaics"; computes surface fluxes over each "tile" using distinct parameters of each tile while sharing the same atmospheric forcing; and then averages the computed fluxes from the tiles, weighted by the fractional areas of each "tile."

Topographic index Or "wetness index," = $\ln(a/\tan\beta)$, where a is the specific catchment area, that is, the upstream area above a pixel that drains through the unit contour at the pixel, and $\tan\beta$ is the local surface topographic slope. A greater value corresponds to a pixel of lowland areas that are potentially wetter.

Definition of the Subject and Its Importance

Land covers about 29% of the Earth's surface and represents an important component of the Earth's climate system. The land surface encompasses soil, snow, vegetation, glaciers, urban, lakes and rivers, mountains, etc. Land surface processes refer to various biogeophysical and biogeochemical processes occurring within and over these components and interacting with the atmospheric processes. Land surface models (LSMs) are mathematical descriptions of the land surface processes, with particular attention to momentum, energy, and mass (water, carbon, dust, and other constituents) flux exchanges with its overlying atmosphere; these fluxes are intimately related to their storages in these components. In the context of an earth system model (ESM), LSMs are designed to provide the atmospheric model with lower boundary conditions over global land areas, describing the interactions of the land surface with the atmosphere. In addition, LSMs provide ocean models with discharges of water, sediments, and solutes. In this chapter, the focus is on descriptions of processes related to momentum, energy, water, and carbon, while processes related to other mass constituents, like dust, sediment, solute transports will not be included. Readers can also further refer to reviews of land surface modeling such as Sellers et al. [1], Pitman [2], and Yang [3].

Most LSMs, if not all, for use in ESMs and operational weather forecasting models are vertically one-dimensional (1D) models solving soil temperature and moisture equations, while parameterizing the effects of lateral flows of surface and subsurface water driven by subgrid topography. Although three-dimensional (3D) hydrological models numerically describing lateral flows of surface and subsurface water and operating at a resolution of meters have existed for years, it is not practical to include them in a climate model due to their expensive computational cost. Vertically, an LSM extends down to a few meters of the soil (2–6 m in different models) and a few meters above the vegetation canopy or the urban canopy. Some LSMs extend their depth to tens of meters deep to consider the long-term effects of permafrost and unconfined aquifers [4].

LSMs describe processes at various spatial scales ranging from millimeter to continental scales. The energy flux exchanges between land surface and the

atmosphere involve processes at various scales from scattering of photons by plants and latent and sensible heat transfer by surface layer small eddies to convective planetary boundary layer large eddies. The water flux exchange involves processes at scales from water infiltration and percolation through soil micropores, transpiration through plant stomata, hillslope overland flow to continental-scale hydrological processes. The subgrid, unresolvable small-scale processes are highly parameterized using empirical equations based on a basic understanding of the processes, but not attempting to represent the process by equations representing first principles. This is why LSMs are also referred to as land surface parameterization schemes.

The land surface processes can influence weather and climate at timescales ranging from seconds to millions of years. For instance, the convective atmospheric boundary layer is developed in response to land surface heating on timescales from seconds to hours. Atmospheric temperature and precipitation can be affected through changes in soil water and groundwater, seasonal variations in snow cover and frozen soil, and vegetation phenology on a scale ranging from days to seasons. Temperature and water availability can greatly influence vegetation structure and function from years to centuries [5]. The Earth's climate has been tightly coupled to atmospheric CO₂ levels through the carbonate silicate cycle and/or the organic carbon cycle on geological timescales spanning hundreds of millions of years [6].

Land surface processes are critical for understanding climate dynamics owing to its momentum, energy, and mass exchanges with the overlying atmosphere. Land surface can “memorize” climate dynamics by recording and filtering the signals of weather events through storages of energy, water, and carbon and feed back to the atmosphere through surface fluxes. Pitman [2] reviewed coupled model experiments including sensitivity studies of changing albedo (e.g., [7]), roughness length (e.g., [8]), LAI [9], water-holding capacity [10], and roots [11], etc., deforestation [12], desertification [13], land cover change (e.g., [14]), ecosystem feedback under doubling CO₂ [15], and vegetation-snow-albedo feedback [16]. These sensitivity experiments provided evidence that climate models are sensitive to the land surface processes [2].

In addition, land surface feedbacks may amplify and prolong sea surface temperature (SST) anomaly triggered droughts, playing a critical role in the persistence of droughts and mega-droughts [17, 18]. It is well known that global climate dynamics is greatly affected by ocean–atmosphere interactions. However, land surface processes are of great importance for seasonal, interannual, and decadal climate change over specific land areas. Koster et al. [19] revealed that land and ocean processes have essentially different geographic regions of influence on precipitation variance, that is, the ocean’s influence is dominant mainly in the tropical land, while the land’s influence is dominant over mid-latitude land areas. Koster et al. [20] identified the arid-to-wet transition regions over land as “hot spots” where changes in surface water fluxes associated with soil water storage have greater impacts on precipitation.

The land surface is more heterogeneous than other Earth surface components. Within one grid-cell of an ESM on the order of ~ 100 km, different land surface types, for example, bare soil, snow, grass, crop, and forests, may coexist. An LSM has to parameterize the integral effects of these subgrid surfaces on surface energy, water, and carbon fluxes over the grid cell. A variety of methods are used, for example, effective parameters, “mosaic” (or tile) methods, probability density function (pdf) methods, or finer grids than the host climate model. LSMs are usually community efforts, because they are inherently interdisciplinary including atmospheric sciences, hydrology, ecology, soil and snow physics, etc., as well as social sciences related to land use and water use managements.

Introduction

LSMs have evolved for three generations [1, 2] since the advent of the first generation global climate models (GCMs). The first LSM was the “bucket model,” which describes the land surface as a bucket with a fixed water holding capacity of 15 cm [21]. The bucket can be filled with precipitation and emptied by evaporation. Runoff is generated as the water that exceeds the bucket holding capacity, and evaporation is limited by the water content of the bucket. In the first generation models, the transfer of sensible and latent heat was based on simple aerodynamic bulk transfer

formulations, using uniform and prescribed surface parameters, for example, water holding capacity, albedo, and roughness length. These models did not account for ecological processes (e.g., transpiration through plants’ stomata) and more detailed hydrological processes (e.g., infiltration and capillary upward flow through soil micropores). Despite their simplifications, the first generation models represent a key step to describe land surface processes in ESMs, and their simulated soil moisture and evapotranspiration (ET) were comparable to more complex models at longer timescales.

From the 1980s to early 1990s, a second generation of models appeared. They explicitly represented vegetation effects and more complex soil hydrology in the calculation of surface energy and water balances [22–24]. Representative second-generation models are the Biosphere-Atmosphere Transfer Scheme (BATS) [23] and the Simple Biosphere Model (SiB) [24]. During this period, two dozen second-generation models were developed with emphasis on various specific processes (e.g., [25–28]). These models vary in detail, but they have many common features. They usually have a number of soil layers, a single canopy layer (or an additional understory canopy layer like in SiB), and a bulk snow layer. Micrometeorological, hydrological, and ecological processes are more explicitly represented in these models. Instead of the fixed surface albedo in the first-generation models, more explicit radiation transfer through the vegetation canopy differentiating visible and near-infrared and beam and diffusive lights was taken into account. Turbulent transfer of heat from multiple surface sources, for example, canopy leaves and the ground surface was considered, and the under-canopy and leaf boundary layer turbulences were parameterized based on atmospheric boundary layer similarity theories. Transpiration through plants’ stomata were considered using the Jarvis-type stomatal conductance scheme [29], a simple empirical formulation as a function of light (photosynthetically active radiation; PAR), water (leaf water potential), and environmental conditions (temperature and humidity). Whether the soil moisture Richards equation is solved or not, vertical gravity and capillary flows of water were taken into account. Runoff was parameterized as various functions of soil water storage or excess above infiltration, neglecting topographic effects.

Accumulation and ablation of snow on the ground were represented with a bulk layer of snow, neglecting snowpack internal processes.

Since the early 1990s, third-generation models have been developed. The third-generation models were mainly characterized by the capability of simulating carbon uptake by plants and vegetation dynamics, or “greening” of the land surface [1, 2]. Other aspects of the model were substantially improved at the same time. Some models also began to incorporate treatments of nutrient dynamics and biogeography, so that vegetation systems can change location in response to climate shifts. Collatz et al. [30] and Sellers et al. [31] began to integrate stomatal conductance and photosynthesis into LSMs based on the work of Farquhar et al. [32]. A semiempirical model of leaf conductance was proposed [30] based on the understandings of the limitations on carbon assimilation by leaf and maximum use of water by plant [33].

During the same time period, LSMs were substantially improved in other aspects with regard to representations of hydrological processes in ESMs. LSMs started to account for subgrid variability in a more explicit way, for example, the “tile” approach to represent vegetation patchiness [34, 35] and the statistical approach to represent subgrid distribution of soil moisture and runoff generation [36], spatial variability of infiltration [37], topographic control on subgrid soil moisture distribution and runoff generation [38, 39], and subgrid snow distributions [40, 41]. LSMs started to include more physically based, multilayer snow submodels to accommodate more internal processes by parameterizing growth of grain size and liquid water retention and percolation within snowpack [42–44]. Ice content within soil (frozen soil) was explicitly resolved as a new prognostic variable, and its impacts on runoff and infiltration were investigated [45–47]. Most recently, groundwater processes were implemented in LSMs [48–52], and LSMs entered a new era involving multiple disciplinary sciences.

Major Processes Represented in an LSM

Subgrid Heterogeneity and Surface Data

Within an ESM grid on the order of ~ 100 km, land surface properties, for example, soil wetness, snow, vegetation, and terrain, vary at the scale of a few meters.

Because of the strongly nonlinear dependence of surface fluxes on surface properties and states, for example, the dependence of ET on soil wetness, simple area-averaging approaches used to aggregate surface properties and states is problematic. The use of LSM grids whose resolution is higher than ESMs may partially improve the accuracy of grid cell averaged fluxes [53] but it is still very difficult to explicitly resolve the variability. Therefore, representing subgrid variability within ESMs will always remain a challenge no matter how fine the model resolution is [36, 54]. There are three major approaches to representing the subgrid heterogeneity, that is, effective parameters, “mosaic” or “tile” approach, and statistical approach.

Earlier LSMs used effective parameters to obtain the correct fluxes averaged over patches of a grid cell. Some effective parameters can be obtained by simply averaging the parameters for various patches, e.g., albedo, because of the linear relationship between radiative fluxes and surface albedo. But other effective parameters, which have a nonlinear relationship with surface fluxes, for example, the effective roughness length, cannot be obtained by simply averaging parameter values of individual patches and need to be derived in such a way that correct area-averaged fluxes [55–57] can be ensured.

The statistical approach uses pdf of input variables to derive grid cell averaged runoff and ET. Assuming point precipitation intensity and soil wetness are spatially distributed over a large ESM grid cell according to exponential and gamma distributions respectively, Entakhabi and Eagleson [36] derived grid cell mean surface runoff ratio (to grid cell mean precipitation), bare soil evaporation efficiency (ratio of actual to potential), and transpiration efficiency (ratio of actual to potential evaporation). Liang et al. [37] implemented spatially variability of infiltration capacity in the Variable Infiltration Capacity (VIC) model. Niu et al. [39] used an exponential distribution of topographic index (or wetness index) to parameterize fractional saturated area and saturation excess runoff.

The tile approach [34, 35] gathers similar homogeneous land patches at different geographic locations within an ESM grid cell into a number of “tiles” or “mosaics,” computes surface fluxes over each “tile” using distinct parameters of each tile while sharing the same atmospheric forcing, and then averages the

computed fluxes from the tiles, weighted by the fractional areas of each “tile.” The performance of the tile approach has been evaluated in numerous offline studies (e.g., [35]), boundary layer or mesoscale atmospheric models (e.g., [34]), and coupled ESMs [26, 58]. The tile can be also generated using different variables, for example, topographic index [38], soil wetness [54], or soil ice content [59]. The tile approach may overlap too much canopy shadows and thus induce errors in computing net solar radiation absorbed by the ground, especially over high latitude, boreal forest regions [60]. However, this approach is widely used in numerous second- and third-generation LSMs, because of its readiness to make full use of high-resolution satellite remotely sensed land surface data to generate the “tiles.”

As an example, the National Center for Atmospheric Research (NCAR) Community Land Model version 4.0 (CLM4) [61] adopted the tile approach and parameterized some key processes using the statistical approach. CLM4 accounts for the subgrid heterogeneity as a nested subgrid hierarchy of landunits, soil/snow columns, and plant functional types (PFTs). Each grid cell can have a number of different landunits, for example, glacier, wetland, lake, urban, natural vegetation, and managed vegetation (crops). Each landunit can have a number of different columns (e.g., irrigated and rainfed soil), and each column can have multiple PFTs. CLM4 adopted the statistical method of Niu et al. [39] to parameterize surface saturated area and associated runoff. Surface datasets to feed CLM4 include percentages of glacier, lake, wetland, urban, and various PFTs to determine the fractional area of these surface tiles. PFTs are based on MODIS satellite data at 30 s (~1 km) resolution [62]. Prescribed PFT LAI was derived from the MODIS satellite data of Myneni et al. [63]. Prescribed PFT stem area index (SAI) is derived from PFT LAI phenology combined with the method of Zeng et al. [64]. Prescribed canopy top and bottom heights are from Bonan et al. [65]. Soil color dataset (affecting surface albedo) is based on Lawrence and Chase [62]. The IGBP soil dataset (Global Soil Data Task 2000) of 4,931 soil mapping units and their sand and clay content for each soil layer is used to create a mineral soil texture dataset [65] and an organic matter density dataset [66] that vary with depth.

Surface Energy Balance

Over an ESM grid cell, the net radiation (R_n) absorbed by the land surface is balanced by latent (LE), sensible (H), and ground heat fluxes (G):

$$R_n = LE + H + G \quad (1)$$

Note that the chemical energy stored during photosynthesis and released by respiration is not included in the above equation as it amounts to less than 1% of the absorbed insolation [2]. R_n includes net shortwave and net longwave radiation fluxes. LSMs that do not differentiate the vegetation canopy and the ground, for example, Noah [67] use the above equation and compute a single temperature for the land surface. More complex models that have a separate canopy layer from the ground, like versions of SiB and CLM, use two energy balance equations, one for the vegetation canopy and another for the ground surface, to solve their temperatures. The partitioning of available energy (R_n) into latent and sensible heat fluxes represented by Eq. 1 is greatly controlled by the land surface hydrological conditions. For an LSM using the tile subgrid approach, Eq. 1 is applied over each tile.

Net Shortwave Radiation LSMs compute net shortwave radiation over various land surfaces:

$$S_a = S(1 - \alpha_s) \quad (2)$$

where S is the incoming solar radiation from the host ESM, and α_s is the surface albedo, varying with solar zenith angle (SZA), vegetation types, snow cover, and soil moisture conditions. In earlier LSMs, α_s was prescribed for various land surface types. More complex second- and third-generation models used Beer's Law and two-stream approximation scheme [68, 69] to compute radiative transfer through the vegetation canopy, net radiation absorbed by the vegetation canopy and the ground, and the surface albedo.

Beer's law assumes that irradiance decreases exponentially with path length through an absorbing medium. Thus, the radiation reaching the ground, $S_g = Se^{(-\kappa LAI)}$, where κ is a constant attenuation coefficient, and LAI is the effective leaf area index of the canopy. However, Beer's law does not account for the scattering and multiple reflections by the canopy leaves [70].

The two-stream approximation [68, 69] accounts for beam and two main streams (vertical upward and downward) of diffusive radiation over two wavebands (visible and near-infrared). However, it assumes that the canopy leaves are evenly distributed within a grid cell. On the basis of complex radiation transfer models [71], Yang and Friedl [72] proposed a modified two-stream scheme to account for aggregation of the evenly distributed leaves represented by a two-stream scheme into canopy crowns with between-canopy and within-canopy gaps, which vary with radius and thickness of the canopy, tree density (the distance between trunks), and SZA. The scheme needs prescribed optical parameters, for example, reflectivity and absorptivity of the vegetation and ground, for various land surface types. The two-stream approximation also computes the fraction of sunlit and shaded leaves, which are crucial to photosynthesis processes.

Net Longwave Radiation The net longwave radiation (or infrared) is computed as the residual of the incoming ($L_{\downarrow}$) and emitted by the surface:

$$L_a = \varepsilon(L_{\downarrow} - \sigma T_s^4) \quad (3)$$

The upward longwave radiative flux emitted by the surface is based on the Stefan-Boltzmann Law. In the above equation σ is the Stefan-Boltzmann constant, ε is the emissivity of the land surface, and T_s is the surface skin temperature. For a model that has a canopy layer, computation of L_a becomes more complicated due to the interactions between the ground surface and its overlaying canopy [61].

Sensible and Latent Heat Fluxes The turbulent transfer of latent and sensible heat fluxes in most LSMs is based on Monin-Obukhov similarity theory [73, 74]. LSMs compute surface sensible and latent heat fluxes using the bulk transfer scheme:

$$H = \rho C_p C_H u (T_0 - T_a) \quad (4)$$

$$LE = \frac{\rho C_p}{\gamma} C_W u (e_0 - e_a) \quad (5)$$

where ρ is air density, C_p is the specific heat of dry air, γ is the psychrometric constant, C_H and C_W are the turbulent exchange coefficients for heat and water vapor, respectively, T_a and u are the air temperature and wind

speed at the lowest layer of the host ESM, and T_0 and e_0 are the temperature and water vapor pressure at the height of the surface roughness length, z_0 . The first-generation models prescribed C_H and C_W with different values for various land surfaces. The second- and third-generation models calculate C_H and C_W according to Monin-Obukhov similarity theory, accounting for surface stability conditions. For simplicity, $C_H = C_W$ and is assumed to be equivalent to that for momentum C_D in some LSMs. More elaborate schemes differentiate C_H ($=C_W$) from C_D using roughness lengths different for heat z_{0H} and momentum z_0 ; and usually $z_0 > z_{0H}$, because the transfer of momentum is more efficient (through pressure and diffusion) than those of heat and mass (through diffusion only). The value of $\ln(z_0/z_{0H})$ is about 2.0 over a range of vegetation surfaces, and in practical applications, $z_0/z_{0H} = 10$ [75]. In analogy with Ohm's law in electricity, the drag coefficient can be transformed to an aerodynamic resistance:

$$r_{a,H} = r_{a,W} = (C_H u)^{-1} \quad (6)$$

Shuttleworth [75] provided a review on formulations of the aerodynamic resistances for estimating surface ET and stability correction.

Complex models compute sensible and latent heat fluxes from multiple sources of the land surface such as soil, snow, plants' leaves, and stomata using equations similar to Eqs. 4 and 5. Over the soil surface, because of the strong temperature gradient between the soil skin (at 0 m) and z_0 as well as the difficulty in computing the temperature at z_0 , LSMs use the skin temperature T_s to compute sensible and latent heat fluxes for convenience. The skin temperature over these surfaces can be computed iteratively using Eq. 1 till the surface energy is balanced. Because of the use of T_s rather than T_0 in Eqs. 4 and 5, an extra resistance, r_{soil} is needed [76]:

$$H = \rho C_p \frac{(T_s - T_a)}{r_{a,H} + r_{soil}} \quad (7)$$

$$LE = \frac{\rho C_p}{\gamma} \frac{(e^*(T_s) - e_a)}{r_{a,W} + r_{soil}} \quad (8)$$

where $e^*(T_s)$ is the saturated vapor pressure at T_s , and r_{soil} accounts for the resistance to water vapor and heat transfer from the surface soil pore space to z_{0H} (roughness sub-layer) through molecular diffusion. Some

LSMs also added a relative humidity term before $e^*(T_s)$ to estimate the water vapor pressure within surface soil pore space. Sakaguchi and Zeng [77] reviewed numerous soil resistance schemes. This expression for soil surface evaporation is similar to that for transpiration through leaf stomata

$$LE = \frac{\rho C_p (e^*(T_v) - e_a)}{\gamma (r_{a,v} + r_c)} \quad (9)$$

where T_v is the vegetation canopy temperature, $r_{a,v}$ is the leaf boundary layer resistance per unit LAI [73], and r_c is the canopy-scale stomatal resistance that is scaled up from the stomatal-scale stomatal resistance, r_{st} , using LAI: $r_c = r_{st}/LAI$. Most second-generation models implemented the Jarvis-type stomatal resistance [29], which relates r_{st} to multiplicative stress factors:

$$1/r_{st} = g_0 f_s(PAR) f_D(\Delta e) f_T(T_a) f_W(\psi_l) \quad (10)$$

where g_0 is a vegetation type-dependent maximum stomatal conductance and f_s , f_D , f_T , and f_W , all of which range from 0 to 1, are environmental stress factors associated with PAR, water vapor deficit of the surrounding air (Δe), air temperature (T_a), and leaf water potential that relates to root-zone soil moisture, respectively.

Ground Heat Flux The ground heat flux is controlled by the soil thermal diffusion (or conduction) process and can be expressed as

$$G = -\lambda \frac{T_1 - T_s}{\Delta z_1/2} \quad (11)$$

where λ is the soil thermal conductivity, T_1 is the temperature of the first soil layer, and Δz_1 is the layer thickness of the first soil layer. By definition ($G = -\lambda(\partial T/\partial z)_{z=0}$), the smaller Δz_1 , the more accurate G is. In case of soil covered by snow, LSMs that have a bulk snow layer replace $\Delta z_1/2$ in the above equation with $(\Delta z_1/2 + h_{\text{snow}})$, where h_{snow} is the snow depth. Thus, when the snowpack becomes thicker, G becomes more inaccurate. This is one of the major reasons that complex models use a multi-layer structure for the snowpack. G is then used as the upper boundary condition of soil temperature equation.

Heat Storage in the Soil and Snow The soil thermal conduction (or diffusion) equation is written as

$$C \frac{\partial T}{\partial t} = \frac{\partial}{\partial z} (\lambda \frac{\partial T}{\partial z}) \quad (12)$$

where C and λ are the volumetric heat capacity and thermal conductivity, respectively. LSMs solve the above equation using various methods, for example, the more efficient force-restore method [22, 78] as used in BATS and SiB and the finite difference method as used in the Noah and CLM4 models.

The force-restore method is based on an analytical solution of the above equation under two assumptions that the surface forcing is periodic and the soil thermal properties are uniform. It is named after the two terms controlling surface temperature, that is, the forcing term (G in Eq. 11) and a restoring term due to heat diffusion from surface to a deeper soil layer. Recognizing that the two assumptions are not always valid, Dickinson [78] took into consideration the nonuniform thermal properties induced by snow and soil moisture and generalized the force-restore method. However, the surface temperature solved through the force-restore method is not so responsive to variations in surface forcing with a frequency higher than diurnal (e.g., cloud effects on solar radiation) as the iterative method solving the surface energy balance equation (Eq. 1). Therefore, it may cause difficulties in resolving the available energy for snowmelt, which occurs at a sub-daily scale [60].

Current LSMs explicitly solve Eq. 12 using finite difference methods for layered soil and snow. The upper boundary condition is the surface forcing (G), and lower (bottom) boundary condition is either specified as a zero flux as used in CLM or as the long-term averaged surface air temperature at a deep soil layer (e.g., 8 m) as used in Noah [67]. The zero flux condition is more readily satisfied when the bottom layer is deeper. The accuracy of the solution increases as the number of layers increases. Due to the computational cost, the number of layers cannot be very large. CLM has ten soil layers, while many other LSMs have three to five layers. To account for the thermal inertia of deep ground, the number of soil layers of CLM4 is changed from 10 to 15 layers, extending the total depth of the soil column down to 42 m [4].

When covered by snow, the soil temperature is solved together with the temperature of snow of a varying number of layers [43, 60]. The layering of snowpack and the number of layers vary with the total snow depth and is more complex than layering soil (e.g., [43, 61]). For frozen soil, a source/sink term due to phase change, $\rho_{ice} L_f \partial \theta_{ice} / \partial t$, where ρ_{ice} , L_f , θ_{ice} , and t are density of ice, heat for fusion, volumetric ice content, and time, respectively, is added to the right-hand side of Eq. 12. This term accounts for the heat released by freezing liquid water or heat consumed to melt ice. It is usually combined with heat conductivity C and named as “apparent heat capacity,” C_a

$$C_a = C - \rho_{ice} L_f \frac{d\theta_{ice}}{dT} \quad (13)$$

Surface Water Balance

The temporal variation of the water storage (S) in a grid cell of an ESM is controlled by the water incident on land surface, that is, precipitation (P) and the water leaving the grid cell, that is, ET and runoff (R):

$$\frac{dS}{dt} = P - ET - R \quad (14)$$

The above equation neglects the lateral surface and subsurface flows that transport water to adjacent grid cells. The storage term in the above equation includes water (ice and liquid water) intercepted by the canopy, snow on the ground, water in the soil and aquifers. ET includes ground evaporation, canopy evaporation, and transpiration, which can be derived from their corresponding latent heat fluxes by dividing the heat of vaporization, L . For most LSMs, R includes surface runoff and subsurface runoff, which will be discussed later.

Canopy-Intercepted Water Although the amount of the canopy-intercepted water is relatively small compared to other water storages, it plays an important role in contributing to the total ET and altering the canopy reflectance by intercepting snowfall. Most LSMs take into consideration liquid water interception by the vegetation canopy, while few LSMs explicitly consider additional processes of loading and unloading of snowfall [79, 80]. The temporal variation of the liquid water stored on the vegetation canopy, $W_{c,liq}$, is

determined by the rates of interception (R_{intr}), evaporation/dew (E_c), and drip (R_{drip}):

$$\frac{dW_{c,liq}}{dt} = R_{intr} - E_c - R_{drip} \quad (15)$$

The interception rate is parameterized as a function of LAI and SAI. The interception capacity of the vegetation canopy is also related to the subgrid distribution of vegetation and precipitation. Considering only a fraction of the vegetation canopy within the large area of an ESM grid cell receives precipitation, CLM4 introduced a fractional area of precipitation as a limiting factor (0.25) to its interception rate [61]. However, this factor should vary with the distribution of precipitation that relates to the ratio of convective precipitation to total precipitation [39] and should be further investigated using high-resolution precipitation data.

The interception capacity for snow is estimated to be about 50 times greater than for rain [80]. Interception of snowfall by the vegetation canopy significantly reduces snow mass on the forest floor. About 30–40% of the annual snowfall over complete coniferous canopies sublimates from the canopy and thus never reaches the ground [80]. Depending on meteorological conditions, the intercepted snow may fall to the ground, sublimate, melt, or refreeze. Implementing a snowfall interception model and properly representing the sublimation of the canopy-intercepted snow can significantly improve the simulation of snow on the ground [81]. More importantly, properly representing interception of snow can significantly improve the simulation of surface albedo. However, most LSMs do not explicitly distinguish between solid and liquid phases of water on the canopy surface, and they use the same interception capacity for both snowfall and rainfall.

Based on the water stored on the vegetation canopy, a wet fraction of the canopy can be computed. Many LSMs (e.g., BATS and CLM) calculate the wet fraction as the 2/3 power of the ratio of canopy water to its maximum value, following Deardorff [22], while others (e.g., SiB) compute it as the ratio of the canopy water storage to its maximum value when the saturation vapor pressure at leaf temperature is less than the vapor pressure of the surrounding air. The maximum value of the canopy water is parameterized as a function of LAI.

Snow Water Snow cover is an important part of the climate system because of its high albedo, low thermal conductivity, and ability to absorb heat when melting. In addition, the snowmelt water is the primary source of stream flow and groundwater recharge in cold and alpine regions. The impacts of snow-cover processes on global and regional climate have been investigated by numerous researchers (e.g., [82–84]) using earlier GCMs. In addition, numerous sensitivity studies using GCMs found that the removal of all forests north of 45°N would lead to cooling and delay snowmelt because of the increased surface albedo (e.g., [12, 85]), indicating the important role of forest canopies in modifying the surface energy budget and climate over snow covered regions.

Snow covered area exhibits spatial and temporal fluctuations ranging from 7% to 40% seasonally in the Northern Hemisphere [86]. Snow cover extent in the Northern Hemisphere has been decreasing since the mid-1980s in response to a global warming trend [87, 88]. On the other hand, Arctic warming may be accelerated by decreases in snow cover due to the positive snow-albedo feedback. Chapin et al. [89] reported that Arctic summer warming is highly related to an increase in snow-free days and the transition from tundra (short to be easily buried by snow) to forest.

LSMs compute snow water equivalent (SWE), snow depth, and fractional area covered by snow. The representation of snow processes in the first- and second-generation LSMs was relatively simple using a bulk snow layer (e.g., [26, 90, 91]). Third-generation models tend to have multiple snow layers to resolve the internal processes of snowpack such as densification, liquid water flow, and multiphase changes of water [42–44, 92]. The multilayer structure also has a thin surface layer designed to produce a more accurate ground heat flux and snow skin temperature [60].

Bulk-layer models combine the snowpack with the top soil layer as a bulk layer. These models, for instance, BATS and SiB, use the force-restore approach to solve snow temperature. However, the force-restore approach cannot accurately resolve high-frequency variations in snow surface forcing and temperature, and thus induce errors in snowmelt. Most of the multilayer snow models were simplified from schemes with detailed internal processes such as grain-size growth and gravitational flows of liquid water within a

snowpack (e.g., [93, 94]). In a multilayer model like CLM4, the snow mass balance includes ice and liquid water content, but neglects the vapor phase. Densification processes include destructive or equi-temperature metamorphism, compaction due to the weight of the overlying layers of snow mass, and melt metamorphism, following Anderson [93]. Excessive liquid water above the holding capacity (ratio of liquid water volume to the free space of a snow layer) flow down to next layer. The melting (or freezing) energy for a layer is assessed as the energy excess (or deficit) needed to change the temperature of the layer to the freezing point [60].

Researchers also recognized that snowpack modeling for hydrological applications and climate studies needs to account for the radiative effects of the vegetation canopy (e.g., [95]) and the impacts of the interception of snow by the vegetation canopy [80, 96]. In addition, the turbulent transfer of sensible heat flux below the vegetation canopy under a stable condition during melting season significantly affects melting snow on the ground [81]. However, these studies were not fully introduced into ESMs.

Snow depth varies greatly at subgrid-scales because of heterogeneities in land cover, terrain, snow deposition, snowmelt, and meteorological conditions [40]. In LSMs, subgrid snow distributions are represented as snow cover fraction (SCF), that is, the fraction of a grid cell covered by snow, through the relationship between SCF and snow depth. At an ESM grid cell scale, one of the largest uncertainties in modeling snow and its interactions with the atmosphere comes from SCF formulations and their associated parameters. Various SCF formulations as summarized in Liston [40] result in a wide spread of SCFs between models. ESMs vary significantly in simulating SCF, and most of them underestimate SCF [97] because of unrealistic SCF formulations. Most SCF formulations are parameterized as a function of grid cell mean snow depth and the ground roughness length [40, 41, 81, 90]. Some of them also considered the impact of subgrid topography variations on SCF [91]. Using gridded snow depth and SWE reanalysis and the advanced very high resolution radiometer (AVHRR) SCF [87], Niu and Yang [41] found the SCF–snow depth relationship varies with seasons and introduced snow density into an existing SCF formulation [90] to approximate the seasonal variation in the SCF–snow depth relationship.

Soil Water Although soil water accounts for a very small portion ($\sim 1\%$) of the terrestrial water storage, it plays an important role in the climate system through its controlling role on ET. Soil moisture anomalies (i.e., deviation from the mean) can persist from weeks to seasons [98, 99]. Numerous researchers have investigated its impact on climate predictability using GCMs (e.g., [100]). Koster et al. [19] revealed that the land's influence is dominant over midlatitude land areas. In analogy to the oceanic hot spot, for example, the eastern equatorial Pacific, playing a key role in the El Niño–La Niña cycle, Koster et al. [20] found the terrestrial “hot spots,” that is, the arid-to-wet transition regions, where precipitation is affected by soil moisture anomalies during boreal summer. These hot spots were further confirmed by Zeng et al. [101] using an alternative index accounting for the land-precipitation coupling strength. For this reason, a proper initialization of soil moisture can enhance precipitation prediction skill over these hot spots [102–104]. Also, upgrades of LSMs with more elaborate representations of soil moisture can improve climate prediction skill especially during ENSO neutral years when the SST signal is relatively weak [104].

Temporal variation in the total soil moisture of a soil column within a certain depth (e.g., 3.4 m in CLM4) is controlled by infiltration, that is, the residual of precipitation minus runoff and evaporation at the soil surface, root uptake through transpiration, and the water flux at the bottom. The vertical distribution of soil water is controlled by vertical root distribution and soil property that determine gravitational and capillary water fluxes. Evolving from the first-generation “bucket” models, which neglects the capillary water flux, some second- and third-generation models directly solve the 1D, mass-based (θ) Richards equation to calculate the vertical distribution of soil moisture [61, 67]. Modern LSMs also tend to increase the number of soil layers to improve the accuracy of the solution. Various model intercomparison projects indicated that LSMs have a much better ability to simulate the soil moisture anomalies than to simulate the absolute value of soil moisture (e.g., [105, 106]).

The mass-based form (θ) of the Richards equation cannot resolve the soil water distribution at both unsaturated zone and saturated zone (groundwater) as continuously as a pressure-based (ψ) Richards equation.

However, when solved through finite-difference numerical scheme, it conserves mass better than the pressure-based form. Zeng and Decker [107] demonstrated that a previous version of CLM cannot maintain the hydrostatic equilibrium of soil moisture distribution, and this problem cannot be resolved by increasing the vertical resolution of the soil column. In CLM4, they proposed a revised form of the Richard equation by explicitly subtracting the hydrostatic equilibrium pressure profile, $\psi_E(z)$, from the soil water pressure profile, $\psi(z)$.

Frozen Soil Frozen soil has a tremendous impact on soil thermal and hydrological properties, ecosystem diversity and productivity, and greenhouse gas release from the soil. Freezing of soil water delays the winter cooling of the land surface; thawing of the frozen soil delays the summer warming of the land surface. Frozen soil also affects the snowmelt runoff and soil hydrology by reducing the soil permeability. Runoff from the Arctic river systems is about 50% of the net flux of freshwater to the Arctic Ocean [108]. This is a large percentage when compared to the freshwater inputs to the tropical oceans, where freshwater input is dominated by precipitation. Runoff and the degree of freshening can affect ocean salinity, sea ice conditions, and thermohaline circulation.

Earlier LSMs did not explicitly solve soil ice content. These LSMs showed a much larger scatter in the simulated soil moisture and runoff in spring than in other seasons due to different representations of the effects of frozen soil on infiltration [109]. Some of them switched off infiltration for subfreezing temperatures, for example, in simplified SiB [25] and BATS [110]. This treatment failed to produce the spring peaks of soil moisture due to the underestimated infiltration [111, 112]. Xue et al. [112] improved SSiB's ability to simulate the spring peaks of soil moisture by gradually decreasing the hydraulic conductivity at a rate of 10% per degree for subfreezing temperatures following SiB2 [113]. Pitman et al. [45] implemented an explicit representation of the hydrological and thermal effects of soil ice in their LSM but found that the representation degraded runoff simulation in a large-scale river basin.

Field studies showed that the effect of frozen soil on infiltration depends on scale, surface conditions of

snow and vegetation, and soil structures. Shanley and Chalmers [114] showed that the effects of frozen soil on runoff are scale-dependent. There was no significant correlation between seasonal runoff ratios and ground frost depth for the 15 year of record from the Sleepers River watershed, USA, with an area of 111 km², while the increased runoff due to frozen ground was observed occasionally in its 0.59-km² agricultural subcatchment. Lindstrom et al. [115] also concluded that there were no clear effects of frozen soil on the timing and magnitude of runoff from an analysis of 16-year data in a 0.5-km² watershed in northern Sweden. Researchers have demonstrated that soil structure, air-filled porosity, ice content, and the number of freezing and thawing cycles are the governing factors affecting the infiltration capacity of frozen soil. Even at very local scales, recent laboratory and field studies using dye tracer techniques (e.g., [116]) revealed that water can infiltrate into deeper soil through preferential pathways where air-filled macropores exist at the time of freezing.

One-dimensional numerical models using the fully coupled heat and mass balance equations [46, 59, 117–119] used a variety of ways to parameterize the hydraulic properties of frozen soil. Some of the models [117, 119] assume that the freezing–thawing process is similar to the drying–wetting process with regard to the dependence of the soil matric potential on the liquid water content. This assumption leads to a very low infiltration rate or even upward water movements resulting in ice heave in surface layers. However, other modelers [46, 47, 59, 118] proposed various schemes to compute hydraulic properties as a means to produce greater infiltration rates. Stähli et al. [118] proposed two separate domains for the water infiltration into frozen soil: the low-flow domain where water flows through the liquid water film absorbed by the soil particles and the high-flow domain where water flows through the air-filled macropores. Koren et al. [46] assumed that frozen soil is permeable due to soil structural aggregates, cracks, dead root passages, and worm holes and reduced the effects of frozen soil on runoff production. Cherkauer and Lettenmaier [59] assumed that surface water tends to find areas of higher infiltration capacity as it flows across a frozen surface. They split their model domain into 10 bins with each having different ice content, which was derived from the observed spatial distribution of soil temperature, to

increase the infiltration rate in VIC model. Niu and Yang [47] introduced fractional permeable area, which partitions an ESM grid cell into an impermeable fraction and a permeable fraction, and used the total soil moisture (liquid water and ice) to calculate soil matric potential and hydraulic conductivity over the permeable fraction.

When soil water freezes, the water closest to soil particles remains in liquid form due to the absorptive and capillary forces exerted by soil particles. For this reason, there is a certain amount of liquid water coexisting with ice over a wide range of temperatures below 0°C. This supercooled soil water at subfreezing temperatures can be computed with the freezing point depression equation [46, 47] or parameterized with observational data [120]. Most LSMs, for example, CLM4, consider ice and liquid water content but neglect the water vapor phase. The energy available for phase change between ice and liquid water is assessed in the same way as for snow except that only the excess above the supercooled liquid water can be frozen in the soil. Li et al. [121] provided a thorough summary of frozen soil models with different levels of complexity.

Groundwater Groundwater storage constitutes about 30% of the terrestrial water storage. Groundwater storage shows as large variations as that of soil water at monthly or longer timescales in Illinois [122]. The groundwater level also shows a strong diurnal cycle in aquifers where the water table depth is less than 2 m because of the water uptake by the roots of the above-ground plants [51]. The rise and fall of the water table directly interacts with soil moisture in the vegetation root zone and thus acts as a source term for ET. Groundwater dynamics also control runoff generation, which can further affect the computation of soil moisture and ET in a climate model. Observational data show that runoff is much more related to the water table depth than to precipitation at monthly timescale in Illinois [50].

Earlier LSMs did not explicitly account for groundwater storage and its interactions with soil moisture. When solving the soil moisture equation, these LSMs assumed zero flux or free drainage at the bottom of the soil column as the lower boundary condition. For instance, an earlier version of CLM (version 3.0) used gravitational free drainage as its lower boundary

condition at a depth of 3.4 m. Water drained from the soil bottom under gravity during wet seasons or wet years (parameterized as subsurface runoff) would never come back to the soil column in dry seasons or dry years. In reality, when the soil is drying, the soil can draw water through capillary suction from its underlying aquifer (or saturated layers), which has a longer memory of the past precipitation events than the soil. LSMs implicitly include such a process when the water table is shallow enough to be within the model soil layers. However, when the water table is below the model's soil bottom, LSMs neglect this upward water flow through capillary forces while taking into consideration gravitational free drainage at the model bottom, resulting in unrealistically dry deep soil.

During the past decade, the need for a groundwater component in LSMs has received increasing attention. As a result, a number of researchers have incorporated a groundwater component into LSMs [48–52, 123]. These models, for use in climate studies, can be divided in two categories: 3D models [51, 123] and 1D models [48–50, 52]. Because the horizontal transport of groundwater is more pronounced at a smaller scale, 3D models are more applicable to regional climate and water resource studies, while 1D models are often used in ESMs.

Researchers [48–50] have proposed more realistic solutions to solve the continuum of soil and relatively shallow aquifers at the expense of adding more integration elements, that is, 100 nodes, 50 layers, and 10 layers, respectively, in the above-referenced studies. Even with these additional layers, the models are only applicable for relatively shallow aquifers (<5 m) because of the prescribed depth of the model. Niu et al. [52] added a single integration element below the soil of CLM and parameterized groundwater recharge rate according to Darcy's law to account for the interactions between soil water and groundwater. Zeng and Decker [107] proposed another simple way to represent the groundwater recharge rate using the deviation of soil water pressure from its hydrostatic equilibrium at the soil bottom. All these developments represent a step forward from the current LSMs using gravitational free drainage or zero flux as their lower boundary condition. Further development and evaluation are needed to more realistically represent groundwater in ESMs.

Runoff Runoff is one of the major components of the global water cycle and accounts for about 40% of the precipitation that falls on land. It plays an important role in the global climate system by affecting ET and freshwater inputs to the oceans, which in turn affects the ocean thermohaline circulation. LSMs' runoff formulation directly affects the partitioning of precipitation into runoff and ET and soil moisture.

Runoff is conceptually difficult to represent in ESMs. The environmental factors that control runoff, precipitation, soil moisture, and topography, often vary considerably on local scales. The plethora of conceptualizations for runoff schemes reflects the uncertainties in understanding and representing fundamental runoff processes. Land model intercomparison projects (e.g., [106, 124]) summarized various implementations of runoff schemes ranging from simple bucket models to more sophisticated topography-based runoff models. The partitioning of precipitation into evapotranspiration, surface runoff and subsurface runoff (baseflow) varies widely among these LSMs. Climate models have been adjusted so that the global, multiyear average runoff production is about one third of the average precipitation. Runoff in BATS is divided approximately equally between surface and subsurface runoff to match the early observational estimates [110].

Most LSMs parameterize runoff processes as a sum of surface runoff (fast component) and subsurface runoff or base flow (slow component). The surface runoff is generated by two major mechanisms: Horton runoff (infiltration excess) and Dunne runoff (saturation excess). The former is assessed as the excess of precipitation intensity over soil infiltration capability, while the latter is estimated as the precipitation over saturated and impermeable surfaces. Horton runoff is more difficult to represent in ESMs because the precipitation intensity produced by ESMs is unrealistically small ("drizzle" problem) and thus precipitation needs to be distributed in both space and time to mimic storms. Hydrologists introduced the concept of fractional saturated area as the dominant control on surface runoff. Surface runoff is largely dominated by the fractional saturated area, that is, the saturated fraction of a grid cell of ESMs, where rainfall or snowmelt water immediately flows into local river systems [125–127].

The fractional saturated area is conceptually correlated with near-surface soil moisture as it is represented

in BATS [110]. More recent implementations [38, 39, 127–129] relate the fractional saturated area to the water table depth (or water deficit depth) and the subgrid characteristics of the topography following TOPMODEL [126, 130]. The latter is described by statistics of the “topographic index” (or “wetness index,” $\lambda = \ln(a/\tan\beta)$, where a is the specific catchment area, i.e., the upstream area above a pixel that drains through the unit contour at the pixel, and $\tan\beta$ is the local surface topographic slope) computed from the high-resolution digital elevation model (DEM) [131]. Famiglietti and Wood [38] proposed a discretized framework in which the distribution of the topographic index was disaggregated into a number of bands, each representing a fraction of the watershed with similar water table depth and soil moisture, to parameterize the subgrid variability in soil moisture and runoff. However, its structural conflicts with climate models and its high computation costs impeded its application to ESMs. Most applications (e.g., [127–130]) used a three-parameter gamma distribution function to represent the pdf of the topographic index. By fitting the pdf of the topographic index with an exponential function, Niu et al. [39] simply expressed the saturated fraction as an exponential function of the water table depth.

The slow component was formulated as drainage under gravity in BATS [110] or a downslope lateral drainage in SiB [24]. TOPMODEL formulated its base flow as an exponential function of the soil water storage deficit [126] or the water table depth [130]. The TOPMODEL-based runoff schemes for ESMs also parameterized base flow as an exponential function of the catchment water storage deficit [128] or the water table depth [38, 39, 127, 129]. The exponential relationship between runoff (surface runoff and base flow) and the water table depth was demonstrated by observations [50].

Carbon Budgets and Vegetation Dynamics

The terrestrial ecosystem stores very large amounts of carbon, which is estimated at about 500 Pg of carbon (1 Pg C = 10^{15} g C) in plant biomass and 2,000 Pg of carbon in soil organic matter (top 1 m) [132]. The atmosphere, with a CO₂ concentration of 370 ppmv, contains about 785 Pg C as CO₂ [132]. Terrestrial

ecosystem absorbs atmospheric CO₂ through photosynthesis at a rate of about 120 Pg C per year (gross primary productivity; GPP) and releases about half of that back to the atmosphere as CO₂ by respiration, resulting in net primary production (NPP) at about 60 Pg C per year. At the same time, heterotrophic respiration from soil and fire return about 60 Pg C per year back to atmospheric CO₂, closing the loop. The carbon fluxes between terrestrial ecosystem and the atmosphere have been quite stable for millennia. However, losses of C from land use change have been steadily increasing over the last one-and-half centuries, approaching rates of about 2 Pg C per year, mostly from tropical deforestation. Net cumulative CO₂ emissions from land use change may be close to 200 Pg C during the human history [132], about 25% of the atmospheric CO₂ at the level of 370 ppmv. Therefore, ESMs simulating the evolution of climate for centuries of the past and the future should represent these processes more realistically in their LSMs. ESMs also need to realistically simulate the relative role of land versus ocean in taking up the CO₂ released by anthropogenic sources.

Representing carbon budgets and vegetation dynamics is a major feature of the third-generation LSMs. The third-generation models try to fully represent plant physiology and phenology. Most of these LSMs use a semiempirical model of leaf conductance [1, 30, 33] to link the stomatal resistance with photosynthesis processes:

$$1/r_{st} = m \frac{A_n}{c_a} \frac{e_a}{e_{sat}} P + g_0 \quad (16)$$

where m is an empirical coefficient (~ 9 for most C₃ plants and ~ 4 for C₄ plants; a larger m means the leaf consumes more water to assimilate the same carbon), A_n is the net carbon assimilation rate, c_a and e_a are the CO₂ concentration and vapor pressure at the leaf surface, respectively, e_{sat} is the saturated vapor pressure at the leaf temperature, P is the atmospheric pressure, and g_0 is the minimum stomatal conductance.

The net carbon assimilation is calculated as the minimum of three limiting factors: the Rubisco-limited, light-limited, and export-limited rates of carboxylation rate for C₃ plants and PEP-carboxylase limitation for C₄ grasses based on Collatz et al. [30, 133]. Both Rubisco-limited and export-limited

rates of carboxylation are related to the maximum rate of carboxylation, which varies with temperature, foliage nitrogen, and soil water:

$$V_{\max} = V_{\max 25} \alpha_{v \max}^{\frac{T_v - 25}{10}} f(N) f(T_v) \beta \quad (17)$$

where $V_{\max 25}$ is maximum carboxylation rate at 25°C depending on plant functional types (PFTs) and $\alpha_{v \max}$ is the Q_{10} parameter (i.e., for an increase in temperature by 10°C, $V_{\max}$ would increase by $\alpha_{v \max}$), $f(T_v)$ is a function that mimics thermal breakdown of metabolic processes [30, 32], $f(N)$ is a foliage nitrogen factor and, the β factor is the soil moisture stress factor.

Most third-generation models, for example, SiB2 [113], IBIS [5], MOSES [117], and CLM4 [61] represent plant physiology using the above formulations. To simulate the net carbon budgets from the terrestrial ecosystem, the models also included representations of autotrophic respiration and heterotrophic respiration. For instance, CLM4 represents autotrophic respiration as the sum of maintenance respiration and growth respiration. Maintenance respiration is a function of temperature and tissue nitrogen concentration for live biomass [134], while growth respiration is assumed to be about 0.3 of the total carbon in new growth on a given time step based on construction cost for a range of woody and nonwoody tissues. CLM4 also represents decomposition, mineralization, and immobilization processes of carbon and nitrogen stored in three litter pools, three soil organic matter pools, and a coarse woody debris pool, under the soil moisture and temperature controls [134].

These models also describe plant phenology. For instance, CLM4 predicts LAI, SAI, and vegetation heights, including three distinct phenological types: evergreen, seasonal-deciduous, and stress-deciduous. The seasonal-deciduous phenology algorithm is based on the parameterizations for leaf onset and offset for temperate deciduous broadleaf forest [135]. Initiation of leaf onset is triggered when a common degree-day summation exceeds a critical value, and leaf litterfall is initiated when daylength is shorter than a critical value. The stress-deciduous phenology algorithm is based in part on the grass phenology model proposed by White et al. [135]. The algorithm handles phenology for vegetation types such as grasses and tropical drought-deciduous trees that respond to both cold and drought

stresses, and that can have multiple growing seasons per year. The overall reasonable performance of CLM4 has been documented in Lawrence et al. [4].

Future Directions

LSMs have been developed for more than 3 decades, focusing more on representations of biogeophysical processes, that is, the energy and water processes occurring in various terrestrial components, and are now at their third generation. They are becoming more complex and more realistically process based than earlier generations. The third-generation models already include terrestrial ecosystem dynamics and associated physiological effects on terrestrial carbon budgets. However, processes affecting greenhouse gas (GHG) emissions from various natural and anthropogenic sources over land are still not fully represented in these models. GHG emissions from various land surfaces are intimately coupled with various terrestrial biogeochemical processes. As suggested by Pitman [2], the need for an LSM to capture biogeochemical processes, which are tightly coupled to hydrological processes, is increasing and this provides focuses for future developments.

Numerous researchers have investigated the feedback of terrestrial ecosystem to climate change for the next century using the third-generation models (e.g., [15, 136]). Despite differences in details, these results demonstrate an important but uncertain role of the future ecosystems in the climate system. In these simulations, the terrestrial ecosystem functions as a sink through about 2050 and then turns into a source due to the collapse of the soil-carbon sink, resulting in an acceleration of global warming (1.5 K higher than without the carbon feedback). However, these results are questionable, because the terrestrial ecosystem and related biogeochemical processes are more complex than what are currently represented in these LSMs. For instance, these models do not fully represent key processes for stabilizing soil carbon, the hysteresis relationship between soil respiration and temperature, and the isoprene protective effects on photosynthesis under high temperature and light stress, etc. As an example, CLM4 includes the controlling role of soil temperature and water potential in the decomposition of leaf litter and soil organic carbon, but it lacks a representation of key processes such as biogeochemical weathering,

heteroaggregate formation, and microbial activities. To represent more sensitive links between climate, microbial physiology, and C transformations, the response of microbial enzymes and active fungi to bacteria ratios to climatic variation and plant cover should be included [137, 138]. CLM4 considered emissions of biogenic volatile organic carbon (BVOC) including isoprene and monoterpenes [139] based on the emission model of Guenther et al. [140]. However, CLM4 neglects the allocation of the assimilated carbon through photosynthesis to BVOC emissions versus respiratory CO₂ emissions and CO₂ inhibitive effects on isoprene emissions. As recently reviewed by Vickers et al. [141], a rich literature exists demonstrating the role of isoprene in protecting photosynthesis under high temperature and light stresses. Therefore, third-generation models have recently started to implement photosynthesis-based isoprene emission algorithms [142]. In addition, a growing body of evidence suggests a temperature-respiration relationship that exhibits a diurnal hysteresis pattern [143]. A representative model (based on a Q₁₀ factor) that does not consider hysteresis would overestimate soil CO₂ efflux for the entire growing season [143]. It has also been found that the hysteresis is caused by the imbalance of production and diffusion of CO₂ within soil, and soil moisture controls the transition from the imbalanced hysteresis pattern under wet soil conditions to a balanced pattern without hysteresis under dry soil conditions. As such, Riveros-Iregui et al. [143] suggested that the role of soil moisture in controlling the hysteresis temperature-respiration relationship should be considered in models. In addition, autotrophic respiration from roots and more elaborate representation of vertical root distribution should also receive more attention.

A major indication of the anthropogenic effects on the climate system is the large-scale conversion of natural ecosystems to cropland throughout human history. Associated with this conversion are changes in GHG emissions and related water use. Global agriculture is a major contributor to increasing the GHGs, with ~18% of the total anthropogenic emissions, including 9% of the total CO₂ emissions due to land use (e.g., deforestation), 40% of the total CH₄ emissions due to rice cultivation, and 62% of the total N₂O emissions due to fertilizer application for year 2000 [144]. For irrigation, global cropland consumes about

2,500 km³ of water each year, which represents almost 70% of the global fresh water withdrawal from surface reservoirs and aquifers [145]. LSMs have started to include crop models and irrigation schemes (e.g., CLM4). Fully representing agricultural practices, for example, sowing, dates, cultivars, irrigation, and fertilization and associated biogeochemical processes that affect GHG emissions and water use represents a direction in future LSM development. Crop models developed by the agriculture community and measurements from the Free-air Controlled Enhancement (FACE) [146] over various natural and agricultural ecosystems should be helpful for developing models and test models' response to enhanced temperature and CO₂ concentration.

Almost all LSMs for use in ESMs are 1D, solving the governing equations in the vertical direction. However, water flows also in lateral directions at surface and in subsurface driven by gravity; the lateral water flow controls the formation and persistence of wetlands and governs the dynamics of freshwater inputs to the oceans. There is an increasing need to implement 3D hydrologic model into LSMs for use in ESMs. Methane emission, which is the second largest contributor to the anthropogenic GHG, is highly related to wetland dynamics. Of all the natural and anthropogenic sources of methane, methane emission from wetlands is known to be the single largest [147]. However, most LSMs lack a representation of wetland dynamics, prescribing wetland area as a fixed value. To more realistically represent wetland area, Fan et al. [148] developed an efficient hydrologic framework to simulate wetlands for climate models, involving large-scale groundwater convergence that feeds wetland. The spatial resolution of ESMs is increasing (~10 km) with recent computational advances. Implementing a quasi 3D hydrological model, which describes lateral surface and saturated subsurface (groundwater) flows and operates at a high spatial resolution (1 km), to LSMs will become possible in the near future. The hydrology community recently proposed hyper-resolution (1 km) global land surface modeling for monitoring Earth's terrestrial water resources, studies of biogeochemical processes, and agricultural applications, etc. [149]. Lessons learned from developing such a hyper-resolution model will certainly be beneficial for developing advanced 3D LSMs for use in ESMs.

Acknowledgments

This review was partially supported by DOE (DE-SC0006773), NSF (AGS-0944101, and NASA (NNX09A021G).

Bibliography

Primary Literature

1. Sellers PJ, Dickinson RE, Randall DA, Betts AK, Hall FG, Berry JA, Collatz GJ, Denning AS, Mooney HA, Nobre CA, Sato N, Field CB, Henderson-Sellers A (1997) Modelling the exchanges of energy, water and carbon between continents and the atmosphere. *Science* 275:502–509
2. Pitman AJ (2003) The evolution of, and revolution in, land surface schemes designed for climate models. *Int J Clim* 23:479–510
3. Yang Z-L (2004) Modeling land surface processes in short-term weather and climate studies. In: Zhu X (ed) *Observations, theory, and modeling of atmospheric variability*, vol 3, World scientific series on meteorology of East Asia. World Scientific, Singapore, pp 288–313
4. Lawrence DM, Oleson KW, Flanner M, Thornton PE, Swenson SS, Lawrence PJ, Zeng X, Yang L, Levis S, Sakaguchi K, Bonan GB, Slater AG (2010) Parameterization improvements and functional and structural advances in version 4 of the community land model. *J Adv Model Earth Syst* 3:M03001, 27p. doi:10.1029/2011MS000045
5. Foley JA, Prentice IC, Ramankutty N, Levis S, Pollard D, Sitch S, Haxeltine A (1996) An integrated biosphere model of land surface processes, terrestrial carbon balance, and vegetation dynamics. *Global Biogeochem Cycles* 10:603–628
6. Pielke RA, Avissar R, Raupach M, Dolman AJ, Zeng X, Denning AS (1998) Interactions between the atmosphere and terrestrial ecosystems: influence on weather and climate. *Glob Chang Biol* 4:461–475
7. Charney J, Quirk WJ, Chow S-H, Kornfield J (1977) A comparative study of the effects of albedo change on drought in semi arid regions. *J Atmos Sci* 34:1366–1385
8. Sud YC, Shukla J, Mintz Y (1988) Influence of land surface roughness on atmospheric circulation and precipitation: a sensitivity study with a general circulation model. *J Appl Meteorol* 27:1036–1054
9. Chase TN, Pielke RA, Kittel TGF, Nemani R, Running SW (1996) Sensitivity of a general circulation model to global changes in leaf area index. *J Geophys Res* 101:7393–7408
10. Milly PCD (1997) Sensitivity of greenhouse summer dryness to changes in plant rooting characteristics. *Geophys Res Lett* 24:269–271
11. Zeng X, Dai J, Dickinson RE, Shaikh M (1998) The role of root distribution for climate simulation over land. *Geophys Res Lett* 25:4533–4536
12. Bonan GB, Pollard D, Thompson SL (1992) Effects of boreal forest vegetation on global climate. *Nature* 359:716–718
13. Xue Y (1997) Biosphere feedback on regional climate in tropical North Africa. *Q J R Meteorol Soc* 123:1483–1515
14. Betts RA (2000) Offset of the potential carbon sink from boreal forestation by decreases in surface albedo. *Nature* 408:187–190
15. Cox PM, Betts RA, Jones CD, Spall SA, Totterdell IJ (2000) Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. *Nature* 408:184–187
16. Foley JA, Kutzback JE, Coe MT, Levis S (1994) Feedbacks between climate and boreal forests during the Holocene epoch. *Nature* 371:52–54
17. Hong SY, Kalnay E (2000) Role of sea surface temperature and soil-moisture feedback in the 1998 Oklahoma–Texas drought. *Nature* 408(6814):842–844
18. Schubert SD, Suarez MJ, Pegen PJ et al (2004) On the cause of the 1930s dust bowl. *Science* 303(5665):1855–1859
19. Koster RD, Suarez MJ, Heiser M (2000) Variance and predictability of precipitation at seasonal-to-interannual timescales. *J Hydrometeorol* 1(1):26–46
20. Koster RD et al (2004) Regions of strong coupling between soil moisture and precipitation. *Science* 305:1138–1140. doi:10.1126/science.1100217
21. Manabe S (1969) Climate and the ocean circulation: 1. The atmospheric circulation and the hydrology of the Earth's surface. *Mon Weather Rev* 97:739–805
22. Deardorff JW (1978) Efficient prediction of ground surface temperature and moisture with inclusion of a layer of vegetation. *J Geophys Res* 83:1889–1903
23. Dickinson RE (1984) Modelling evapotranspiration for three dimensional global climate models. In: Hansen JE, Takahashi T (eds) *Climate processes and climate sensitivity*, vol 29, Geophysical monograph. American Geophysical Union, Washington, DC, pp 58–72
24. Sellers PJ, Mintz Y, Sud YC, Dalcher A (1986) A simple biosphere model (SiB) for use within general circulation models. *J Atmos Sci* 43:505–531
25. Xue Y, Sellers PJ, Kinter JL, Shukla J (1991) A simplified biosphere model for climate studies. *J Climate* 4:345–364
26. Verseghy DL, McFarlane NA, Lazare M (1993) CLASS – a Canadian land surface scheme for GCMs. II: vegetation model and coupled runs. *Int J Climatol* 13:347–370
27. Thompson SL, Pollard D (1995) A global climate model (GENESIS) with a land-surface transfer scheme (LSX). Part 1: present climate simulation. *J Climate* 8:732–761
28. Desborough CE, Pitman AJ (1998) The BASE land surface model. *Global Planet Change* 19:3–18
29. Jarvis PG (1976) The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field. *Philos Trans R Soc Lond A Ser B* 273:593–610
30. Collatz GJ, Ball JT, Grivet C, Berry JA (1991) Physiological and environmental regulation of stomatal conductance, photosynthesis and transpiration: a model that includes a laminar boundary layer. *Agr Forest Meteorol* 54:107–136
31. Sellers PJ, Berry JA, Collatz GJ, Field CB, Hall FG (1992) Canopy reflectance, photosynthesis and transpiration. III. A reanalysis

- using improved leaf models and a new canopy integration scheme. *Remote Sens Environ* 42:187–216
32. Farquhar GD, von Caemmerer S, Berry JA (1980) A biochemical model of photosynthetic CO₂ assimilation in leaves of C3 species. *Planta* 149:78–90
 33. Ball JT (1988) An analysis of stomatal conductance. PhD thesis, Stanford University, Stanford
 34. Avissar R, Pielke RA (1989) A parameterization of heterogeneous land surfaces for atmospheric numerical models and its impact on regional meteorology. *Mon Weather Rev* 117:2113–2136
 35. Koster RD, Suarez MJ (1992) Modelling the land surface boundary in climate models as a composite of independent vegetation stands. *J Geophys Res* 97:2697–2715
 36. Entekhabi D, Eagleson PS (1989) Land surface hydrology parameterization for atmospheric general circulation models including subgrid scale spatial variability. *J Climate* 2:816–831
 37. Liang X, Lettenmaier DP, Wood EF, Burges SJ (1994) A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J Geophys Res* 99:415–428
 38. Famiglietti JS, Wood EF (1994) Multi-scale modeling of spatially variable water and energy balance processes. *Water Resour Res* 30(11):3061–3078
 39. Niu G-Y, Yang Z-L, Dickinson RE, Gulden LE (2005) A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models. *J Geophys Res* 110:D21106. doi:10.1029/2005JD006111
 40. Liston GE (2004) Representing subgrid snow cover heterogeneities in regional and global models. *J Climate* 17:1381–1397
 41. Niu G-Y, Yang Z-L (2007) An observation-based formulation of snow cover fraction and its evaluation over large North American river basins. *J Geophys Res* 112:D21101. doi:10.1029/2007JD008674
 42. Jin JM, Gao X, Yang ZL, Bales RC, Sorooshian S, Dickinson RE (1999) Comparative analyses of physically based snowmelt models for climate simulations. *J Climate* 12:2643–2657
 43. Dai Y et al (2003) The common land model (CLM). *Bull Am Meteorol Soc* 84(8):1013–1023
 44. Yang Z-L, Niu G-Y (2003) The versatile integrator of surface and atmosphere processes (VISA) part 1: model description, global planet. *Change* 38:175–189
 45. Pitman AJ, Slater AG, Desborough CE, Zhao M (1999) Uncertainty in the simulation of runoff due to the parameterization of frozen soil moisture using the GSWP methodology. *J Geophys Res* 104:16 879–16 888
 46. Koren V, Schaake JC, Mitchell KE, Duan Q-Y, Chen F, Baker JM (1999) A parameterization of snowpack and frozen ground intended for NCEP weather and climate models. *J Geophys Res* 104:19,569–19,585. doi:10.1029/1999JD900232
 47. Niu G-Y, Yang Z-L (2006) Effects of frozen soil on snowmelt runoff and soil water storage at a continental scale. *J Hydrometeorol* 7:937–952. doi:10.1175/JHM538.1
 48. Liang X, Xie Z, Huang M (2003) A new parameterization for surface and groundwater interactions and its impact on water budgets with the variable infiltration capacity (VIC) land surface model. *J Geophys Res* 108(D16):8613. doi:10.1029/2002JD003090
 49. Maxwell RM, Miller NL (2005) On the development of a coupled land surface and groundwater model. *J Hydrometeorol* 6:233–247
 50. Yeh PJF, Eltahir EAB (2005) Representation of water table dynamics in a land surface scheme, part I: model development. *J Climate* 18(12):1861–1880
 51. Fan Y, Miguez-Macho G, Weaver CP, Walko R, Robock A (2007) Incorporating water table dynamics in climate modeling: 1. Water table observation and equilibrium water table simulations. *J Geophys Res* 112:D10125. doi:10.1029/2006JD008111
 52. Niu G-Y, Yang Z-L, Dickinson RE, Gulden LE, Su H (2007) Development of a simple groundwater model for use in climate models and evaluation with gravity recovery and climate experiment data. *J Geophys Res* 112:D07103. doi:10.1029/2006JD007522
 53. Hahmann AN, Dickinson RE (2001) A fine-mesh land approach for general circulation models and its impact on regional climate. *J Climate* 14:1634–1646
 54. Sellers PJ, Fennessy MJ, Dickinson RE (2007) A numerical approach to calculating soil wetness and evapotranspiration over large grid areas. *J Geophys Res* 112:D18106. doi:10.1029/2007JD008781
 55. Mahrt L (1987) Grid-averaged surface fluxes. *Mon Weather Rev* 115:1550–1560
 56. Mason PJ (1988) The formation of areally-averaged roughness lengths. *Q J R Meteorol Soc* 114:399–420
 57. Garratt JR (1992) The atmospheric boundary layer. Cambridge University Press, Cambridge
 58. Essery RLH, Besst MJ, Betts RA, Cox PM (2003) Explicit representation of subgrid heterogeneity in a GCM land surface scheme. *J Hydrol* 4:530–543
 59. Cherkauer K, Lettenmaier DP (2003) Simulation of spatial variability in snow and frozen soil. *J Geophys Res* 108:8858. doi:10.1029/2003JD003575
 60. Niu G-Y et al (2011) The community Noah land surface model with multi-parameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J Geophys Res* 116:D12109. doi:10.1029/2010JD015139
 61. Oleson KW et al (2010) Technical description of version 4.0 of the community land model. NCAR Technical Note NCAR/TN-478+STR. National Center for Atmospheric Research, Boulder, 257 pp
 62. Lawrence PJ, Chase TN (2007) Representing a MODIS consistent land surface in the community land model (CLM 3.0). *J Geophys Res* 112:G01023. doi:10.1029/2006JG000168
 63. Myneni RB et al (2002) Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. *Remote Sens Environ* 83:214–231
 64. Zeng X, Shaikh M, Dai Y, Dickinson RE, Myneni R (2002) Coupling of the common land model to the NCAR community climate model. *J Climate* 15:1832–1854

65. Bonan GB, Levis S, Kergoat L, Oleson KW (2002) Landscapes as patches of plant functional types: an integrating concept for climate and ecosystem models. *Global Biogeochem Cycles* 16:5.1–5.23
66. Lawrence DM, Slater AG (2008) Incorporating organic soil into a global climate model. *Clim Dyn* 30. doi:10.1007/s00382-007-0278-1
67. Ek MB, Mitchell KE, Lin Y, Rogers E, Grunmann P, Koren V, Gayno G, Tarpley JD (2003) Implementation of Noah land surface model advancements in the national centers for environmental prediction operational mesoscale Eta model. *J Geophys Res* 108(D22):8851. doi:10.1029/2002JD003296
68. Dickinson RE (1983) Land surface processes and climate-surface albedos and energy balance. In: Saltzman B (ed) *Theory of climate*, vol 25, *Advances in geophysics*. Academic, San Diego, pp 305–353
69. Sellers PJ (1985) Canopy reflectance, photosynthesis and transpiration. *Int J Remote Sens* 6:1335–1372. doi:10.1080/01431168508948283
70. Nijssen B, Lettenmaier D (1999) A simplified approach for predicting shortwave radiation transfer through boreal forest canopies. *J Geophys Res* 104(D22):27,859–27,868
71. Li X, Strahler AH, Woodcock CR (1995) A hybrid geometric optical and radiative transfer approach for modeling albedo and directional reflectance of discontinuous canopies. *IEEE Trans Geosci Remote Sens* 33(2):466–480
72. Yang R, Friedl M, Ni W (2001) Parameterization of shortwave radiation fluxes for nonuniform vegetation canopies in land surface models. *J Geophys Res* 106:14,275–14,286
73. Brutsaert WA (1982) *Evaporation into the atmosphere*. Reidel, Dordrecht, 299 pp
74. Garratt JR, Francey RJ (1978) Bulk characteristics of heat transfer for the instable baroclinic atmospheric boundary layer. *Bound -Lay Meteorol* 15:399–421
75. Shuttleworth WJ (2007) Putting the 'vap' into evaporation. *Hydrol Earth Syst Sci* 11(1):210–244
76. Zeng X, Dickinson RE (1998) Effect of surface sublayer on surface skin temperature and fluxes. *J Climate* 11:537–550
77. Sakaguchi K, Zeng X (2009) Effects of soil wetness, plant litter, and under-canopy atmospheric stability on ground evaporation in the community land model (CLM3.5). *J Geophys Res* 114:D01107. doi:10.1029/2008JD010834
78. Dickinson RE (1988) The force-restore model for surface temperatures and its generalizations. *J Climate* 1:1086–1097
79. Roesch A, Wild M, Gilgen H, Ohmura A (2001) A new snow cover fraction parameterization for ECHAM4 GCM. *Clim Dyn* 17:933–946
80. Hedstrom NR, Pomeroy JW (1998) Measurements and modeling of snow interception in the boreal forest. *Hydrol Proc* 12:1611–1625
81. Niu G-Y, Yang Z-L (2004) Effects of vegetation canopy processes on snow surface energy and mass balances. *J Geophys Res* 109:D23111. doi:10.1029/2004JD004884
82. Yeh T-C, Wetherald RT, Manabe S (1983) A model study of the short-term climate and hydrologic effects of sudden snow-cover removal. *Mon Weather Rev* 111:1013–1024
83. Walsh JE, Ross B (1988) Sensitivity of 30-day dynamical forecast to continental snow cover. *J Climate* 1:739–754
84. Cess RD et al (1991) Intercomparison of snow-feedback as produced by general circulation models. *Science* 253: 888–892
85. Douville H, Royer JF (1997) Influence of the temperate and boreal forests on the northern hemisphere climate in the meteo-France climate model. *Clim Dyn* 13:57–74
86. Hall DK (1988) Assessment of polar climate change using satellite technology. *Rev Geophys* 26:26–39
87. Robinson DA, Frei A (2000) Seasonal variability of northern hemisphere snow extent using visible satellite data. *Prof Geogr* 52(2):307–315
88. Brown RD (2000) North hemisphere snow cover variability and change, 1915–97. *J Climate* 13:2,339–2,355
89. Chapin FS, Sturm M, Serreze MC et al (2005) Role of land-surface changes in Arctic summer warming. *Science* 310:657–660
90. Yang Z-L, Dickinson RE, Robock A, Vinnikov KY (1997) Validation of the snow sub-model of the biosphere-atmosphere transfer scheme with Russian snow cover and meteorological observational data. *J Climate* 10:353–373
91. Douville H, Royer J-H, Mahfouf J-F (1995) A new snow parameterization for the meteo-France climate model, part I: validation in standalone experiment. *Clim Dyn* 12:21–35
92. Lynch-Stieglitz M (1994) The development and validation of a simple snow model for the GISS GCM. *J Climate* 7: 1842–1855
93. Anderson EA (1976) A point energy and mass balance model of a snow cover. NOAA Technical Report NWS 19. Office of Hydrology, National Weather Service, Silver Spring, 150 pp
94. Jordan R (1991) A one-dimensional temperature model for a snow cover. Special Report 91–16. Cold Regions Research and Engineering Laboratory, U.S. Army Corps of Engineers, Hanover
95. Davis RE, Hardy JP, Ni W, Woodcock C, McKenzie JC, Jordan R, Li X (1997) Variation of snow cover ablation in the boreal forest: a sensitivity study on the effects of conifer canopy. *J Geophys Res* 102:29,389–29,385
96. Essery R, Pomeroy J, Parviainen J, Storck P (2003) Sublimation of snow from coniferous forests in a climate model. *J Climate* 16(11):1855–1864
97. Frei A, Gong G (2005) Decadal to century scale trends in North America snow extent in coupled atmosphere–ocean general circulation models. *Geophys Res Lett* 32:L18502. doi:10.1029/2005GL023394
98. Pielke RA, Liston GE, Eastman JL, Lu L, Coughenour M (1999) Seasonal weather prediction as an initial value problem. *J Geophys Res* 104:19,463–19,479. doi:10.1029/1999JD900231

99. Schlosser CA, Milly PCD (2002) A model-based investigation of soil moisture predictability and associated climate predictability. *J Hydrometeorol* 3:483–501. doi:10.1175/1525-7541(2002)003<0483: AMBIO>2.0.CO;2
100. Shukla J, Mintz Y (1982) The influence of land-surface evapotranspiration on the Earth's climate. *Science* 215(4539):1498–1501
101. Zeng X, Barlage M, Castro C, Fling K (2010) Comparison of land-precipitation coupling strength using observations and models. *J Hydrometeorol* 11:980–995
102. Koster RD et al (2004) Realistic initialization of land surface states: impacts on subseasonal forecast skill. *J Hydrometeorol* 5:1049–1063. doi:10.1175/JHM-387.1
103. Koster RD et al (2010) Contribution of land surface initialization to subseasonal forecast skill: first results from a multi-model experiment. *Geophys Res Lett* 37:L02402. doi:10.1029/2009GL041677
104. Yang RQ, Mitchell K, Meng J, Ek M (2011) Summer-season forecast experiments with the NCEP climate forecast system using different land models and different initial land states. *J Climate* 24:2319–2334. doi:10.1175/2010JCLI3797.1
105. Entin JK, Robock A, Vinnikov KY, Zabelin V, Liu S, Namkhai A, Adyasuren T (1999) Evaluation of global soil wetness project soil moisture simulations. *J Meteorol Soc Jpn* 77:183–198
106. Boone A et al (2004) The Rhone-aggregation land surface scheme intercomparison project: an overview. *J Climate* 17:187–208
107. Zeng X, Decker M (2009) Improving the numerical solution of soil moisture-based Richards equation for land models with a deep or shallow water table. *J Hydrometeorol* 10:308–319. doi:10.1175/2008JHM1011.1
108. Barry RG, Serreze MC (2000) Atmospheric components of the Arctic Ocean freshwater balance and their interannual variability. In: Lewis EL et al (eds) *The freshwater budget of the Arctic Ocean*. Kluwer Academic Publishers, New York, pp 45–56
109. Luo LF et al (2003) Effects of frozen soil on soil temperature, spring infiltration, and runoff: results from the PILPS 2(d) experiment at Valdai, Russia. *J Hydrometeorol* 4:334–351
110. Dickinson RE, Henderson-Sellers A, Kennedy PJ (1993) Biosphere-atmosphere transfer scheme (BATS) version 1e as coupled to the NCAR community climate model. NCAR Technical Note, NCAR/TN- 387 + STR. National Center for Atmospheric Research, Boulder
111. Robock A, Vinnikov KY, Schlosser CA, Speranskaya NA, Xue Y (1995) Use of midlatitude soil moisture and meteorological observations to validate soil moisture simulations with biosphere and bucket models. *J Climate* 8:15–35
112. Xue Y, Zeng FJ, Schlosser CA (1996) SSiB and its sensitivity to soil properties: a case study using HAPEX-mobilhy data. *Global Planet Change* 13:183–194
113. Sellers PJ, Randall DA, Collatz GJ, Berry JA, Field CB, Dazlich DA, Zhang C, Collelo GD, Bounoua L (1996) A revised land surface parameterization (SiB2) for atmospheric GCMs: part I. Model formulation. *J Climate* 9:676–705. doi:10.1175/1520-0442(1996)009<0676:ARLSPF>2.0.CO;2
114. Shanley JB, Chalmers A (1999) The effect of frozen soil on snowmelt runoff at Sleepers River, Vermont. *Hydrol Process* 13:1843–1857
115. Lindstrom G, Bishop K, Lofvenius MO (2002) Soil frost and runoff at Svartberget, northern Sweden – measurements and model analysis. *Hydrol Process* 16:3379–3392
116. Stadler D, Stähli HM, Aeby P, Flühler H (2000) Dye tracing and image analysis for quantifying water infiltration into frozen soils. *Soil Sci Soc Am J* 64:505–516
117. Cox PM, Betts RA, Bunton CB, Essery RLH, Rowntree PR, Smith J (1999) The impact of new land surface physics on the GCM simulation of climate and climate sensitivity. *Clim Dyn* 15:183–203
118. Stähli M, Nyberg L, Mellander P-E, Jansson P-E, Bishop KH (2001) Soil frost effects on soil water and runoff dynamics along a boreal forest transect: 2. Simulations. *Hydrol Process* 15:927–941
119. Hansson K, Simunek J, Mizoguchi M, Lundin L-C, van Genuchten MT (2004) Water flow and heat transport in frozen soil: numerical solution and freeze-thaw applications. *Vadose Zone J* 3:693–704
120. Decker M, Zeng X (2006) An empirical formulation of soil ice fraction based on in situ observations. *Geophys Res Lett* 33:L05402. doi:10.1029/2005GL024914
121. Li Q, Sun S, Xue Y (2010) Analyses and development of a hierarchy of frozen soil models for cold region study. *J Geophys Res* 115:D03107. doi:10.1029/2009JD012530
122. Rodell M, Famiglietti JS (2001) An analysis of terrestrial water storage variations in Illinois with implications for the gravity recovery and climate experiment (GRACE). *Water Resour Res* 37(5):1327–1339
123. Kollet SJ, Maxwell RM (2008) Capturing the influence of groundwater dynamics on land surface processes using an integrated, distributed watershed model. *Water Resour Res* 44:W02402. doi:10.1029/2007WR006004
124. Bowling LC et al (2003) Simulation of high latitude hydrological processes in the torne-kalix basin: PILPS phase 2e. 1 – Experimental description and summary intercomparisons. *Global Planet Change* 38:1–30
125. Dunne T, Black RD (1970) Partial area contributions to storm runoff in a small New England watershed. *Water Resour Res* 6(5):1296–1311
126. Beven KJ, Kirkby MJ (1979) A physically based, variable contributing model of basin hydrology. *Hydrol Sci Bull* 24:43–69
127. Stieglitz M, Rind D, Famiglietti J, Rosenzweig C (1997) An efficient approach to modeling the topographic control of surface hydrology for regional and global modeling. *J Climate* 10:118–137
128. Ducharme A, Koster RD, Suarez MJ, Stieglitz M, Kumar P (2000) A catchment-based approach to modeling land surface

- processes in a general circulation model: 2. Parameter estimation and model demonstration. *J Geophys Res* 105(D20):24,823–24,838
129. Chen J, Kumar P (2001) Topographic influence on the seasonal and inter-annual variation of water and energy balance of basins in North America. *J Climate* 14:1989–2014
 130. Sivapalan M, Beven K, Wood EF (1987) On hydrologic similarity: 2. A scaled model of storm runoff production. *Water Resour Res* 23:2266–2278
 131. Quinn PF, Beven J, Lamb R (1995) The $\ln(a/\tan b)$ index: how to calculate it and how to use it within the TOPMODEL framework. *Hydrol Process* 9:161–182
 132. Janzen HH (2004) Carbon cycling in earth systems – a soil science perspective. *Agric Ecosyst Environ* 104(3):399–417. doi:10.1016/j.agee.2004.01.040
 133. Collatz GJ, Ribas-Carbo M, Berry JA (1992) A coupled photosynthesis–stomatal conductance model for leaves of C4 plants. *Aust J Plant Physiol* 19:519–538
 134. Thornton PE, Rosenbloom NA (2005) Ecosystem model spin-up: estimating steady state conditions in a coupled terrestrial carbon and nitrogen cycle model. *Ecol Model* 189:25–48
 135. White MA, Thornton PE, Running SW (1997) A continental phenology model for monitoring vegetation responses to interannual climatic variability. *Global Biogeochem Cycle* 11:217–234
 136. Cramer W, Bondeau A, Woodward FI, Prentice IC, Betts RA, Brovkin V, Cox PM, Fisher V, Foley JA, Friend AD, Kucharik C, Lomas MR, Ramankutty N, Sitch S, Smith B, White A, Young-Molling C (2001) Global response of terrestrial ecosystem structure and function to CO₂ and climate change: results from six dynamic global vegetation models. *Glob Chang Biol* 7:357–373
 137. Schimel JP, Weintraub MN (2003) The implications of exoenzyme activity on microbial carbon and nitrogen limitation in soil: a theoretical model. *Soil Biol Biochem* 35:549–563
 138. Allison SD, Wallenstein MD, Bradford MA (2010) Soil-carbon response to warming dependent on microbial physiology. *Nat Geosci*. doi:10.1038/NGEO846 (Advance online publication)
 139. Levis S, Wiedinmyer C, Bonan GB, Guenther A (2003) Simulating biogenic volatile organic compound emissions in the community climate system model. *J Geophys Res* 108(D21):4659. doi:10.1029/2002JD003203
 140. Guenther A, Karl T, Harley P, Wiedinmyer C, Palmer PI, Geron C (2006) Estimates of global terrestrial isoprene emissions using MEGAN (model of emissions of gases and aerosols from nature). *Atmos Chem Phys* 6:3181–3210
 141. Vickers CE, Gershenzon J, Lerdau MT, Loreto F (2009) A unified mechanism of action for volatile isoprenoids in plant abiotic stress. *Nat Chem Biol* 5:283–291
 142. Pacifico F, Harrison SP, Jones CD, Arneth A, Sitch S, Weedon GP, Barkley MP, Palmer PI, Serca D, Potosnak M, Fu TM, Goldstein A, Bai J, Schurgers G (2011) Evaluation of a photosynthesis-based biogenic isoprene emission scheme in JULES and simulation of isoprene emissions under present-day climate conditions. *Atmos Chem Phys* 11:4371–4389
 143. Riveros-Iregui DA, Emanuel RE, Muth DJ, McGlynn BL, Epstein HE, Welsch DL, Pacific VJ, Wraith JM (2007) Diurnal hysteresis between soil CO₂ and soil temperature is controlled by soil water content. *Geophys Res Lett* 34:L17404. doi:10.1029/2007GL030938
 144. Olivier JGJ, Berdowski JJM (2001) Global emissions sources and sinks. In: Berdowski J, Guicherit R, Heij BJ (eds) *The climate system*. A.A. Balkema/Swets & Zeitlinger, Lisse, pp 33–78. ISBN 90 5809 255 0
 145. Rost S, Gerten D, Bondeau A, Lucht W, Rohrer J, Schaphoff S (2008) Agricultural green and blue water consumption and its influence on the global water system. *Water Resour Res* 44:W09405. doi:10.1029/2007WR006331
 146. Kimball BA (2011) Lessons from FACE: CO₂ effects and interactions with water, nitrogen, and temperature. In: Hillel D, Rosenzweig C (eds) *Handbook of climate change and agroecosystems: impacts, adaptation, and mitigation*. World Scientific, Singapore, pp 81–107
 147. Houghton JT et al (eds) (2001) *Projections of future climate change, climate change 2001: the scientific basis, contribution of working group I to the third assessment report of the intergovernmental panel on climate change*. Cambridge University Press, New York, 881 pp
 148. Fan Y, Miguez-Macho G (2011) A simple hydrologic framework for simulating wetlands in climate and earth system models. *Clim Dyn* 37:253–278. doi:10.1007/s00382-010-0829-8
 149. Wood EF et al (2011) Hyperresolution global land surface modeling: meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resour Res* 47:W05301. doi:10.1029/2010WR010090

Earth System Monitoring, Introduction

JOHN ORCUTT

Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA

This section provides sensing and data collection methodologies, as well as an understanding of Earth's climate parameters and natural and man-made phenomena, to support a scientific assessment of the Earth system as a whole, and its response to natural and human-induced changes. The coverage ranges from climate change factors and extreme weather and fires to oil spill tracking and volcanic eruptions. This serves as a basis to enable improved prediction and response to

climate change, weather, and natural hazards as well as dissemination of the data and conclusions. The data collection systems include satellite remote sensing, aerial surveys, and land- and ocean-based monitoring stations. Our objective in this treatise is to provide a significant portion of the scientific and engineering basis of Earth system monitoring and to provide this in 17 detailed articles or chapters written at a level for use by university students through practicing professionals. The reader is also directed to the closely related sections on ► [Ecological Systems, Introduction](#) and also ► [Climate Change Modeling Methodology, Introduction](#) as well as ► [Climate Change Remediation, Introduction](#). For ease of use by students, each article begins with a glossary of terms, while at an average length of 25 print pages each, sufficient detail is presented for use by professionals in government, universities, and industries. The chapters are individually summarized below.

► [Aircraft and Space Atmospheric Measurements Using Differential Absorption Lidar \(DIAL\)](#) – A particular lidar technique called differential absorption lidar (DIAL) has proven to be an especially powerful method to measure atmospheric chemical species. It has been used both on the ground and in many aircraft to determine the concentration of gasses. DIAL systems can be deployed in fixed ground locations, in either piloted or unpiloted aircraft or in satellite platforms.

► [Coral Reef Ecosystems](#) – Over the millennia, coral reef ecosystems have served as the basis of human survival and sustenance in many coastal areas in the tropics and subtropics. They are significant geologic formations that provide crucial protection from destructive forces of the sea. They also harbor the greatest diversity of animal phyla of any other habitat on the planet. Threats due to climate change are detailed.

► [Airborne and Space-borne Remote Sensing of Cryosphere](#) – The cryosphere broadly constitutes all the components of the Earth system, which contain water in a frozen state. As such, glaciers, ice sheets, snow cover, lake and river ice, sea ice, and permafrost make up the terrestrial oceanic elements of the cryosphere. While ice particles in the upper atmosphere and icy precipitation near the surface are the representative members of the Cryosphere in atmospheric systems. Airborne and spaceborne remote sensing technologies are presented.

► [Heat Content and Temperature of the Ocean](#) – Following years of research and development in autonomous float technology, major international cooperation under the Argo Project led to the seeding of the global ocean with more than 3,000 floats each with the capability of profiling temperature and other properties between depths of 2,000 m and the surface approximately every 10 days. Measurements are relayed by satellite while the floats are at the surface in between each cycle of ascent and descent. Coverage is almost global. Measuring variability in sea level using satellite altimetry in situ temperature observations are complementary in efforts to infer changes in heat content.

► [Hurricane and Monsoon Tracking with Driftsondes](#) – During the international African Monsoon Multidisciplinary Analysis (AMMA) project, high flying balloons (called driftsondes) dropping meteorological sondes to take measurements over remote, but scientifically important locations, were deployed over West Africa and tropical Atlantic to study weak weather systems, called African easterly waves, that serve as seedlings for hurricanes.

► [Large-Scale Ocean Circulation: Deep Circulation and Meridional Overturning](#) – Observations from long-term moorings and deep floats have illustrated a dynamic deep ocean environment, with energetic variability observed at a wide range of space and time scales, including narrow jets, eddies, and recirculating gyres.

► [Long-Term Ecological Research Network](#) – The US Long Term Ecological Research (LTER) Network is the largest and longest-lived ecological network in the continental United States. Designed to provide long-term data from a broad range of key ecosystems, the LTER Network represents a unique national resource for addressing pressing environmental issues such as climate change, loss of biodiversity, and changes in patterns of land use.

► [Ocean Acidification](#) – The oceans have absorbed about one third of the anthropogenic carbon emissions, which in part ameliorates global warming induced by human activity. However, this capacity to absorb CO₂ is challenged by the accelerated increase in human activities including the increasing burning of fossil fuels, cement production, agriculture, and deforestation. The accelerated rate of increase in the concentration of atmospheric CO₂ and the parallel rise in dissolved CO₂ in seawater (and the associated chemical

changes) can affect marine life. For example, increases in CO₂ can potentially disrupt the acid base balance in tissues, fluids, and cells of marine organisms.

► **Ocean Evaporation and Precipitation** – Taken as a whole, the hydrological cycle is characterized by the evaporation of about half a million cubic kilometers of water per year, the bulk of which (86%) is from the ocean, with only 14% originating in the continents. The vast majority of the water that evaporates from the oceans (90%) is precipitated back into them, while the remaining 10% is transported to the continents, where it precipitates. About two thirds of this precipitation is recycled over the continents and only one third runs off directly into the oceans. This chapter covers the distribution of evaporation from the ocean and precipitation, oceanic heat reservoirs and evaporation, oceanic evaporation and monsoonal precipitation regimes, and the response of the relationship between evaporation and precipitation.

► **Ocean Observatories and Information: Building a Global Ocean Observing Network** – Ocean observatories are collections of networks of sensors that are deployed to sample the ocean physics, chemistry, and biology. The goal of these networks is to overcome chronic under-sampling of the oceans by providing sustained measurements in space and time. The data collected by these networks are used to address a range of basic and applied research questions, hindered by a lack of data. This chapter covers the need for a global ocean observing network, design considerations, platforms, and information systems.

► **Oil Spill Remote Sensing** – Remote sensing for oil spills is reviewed. The technical aspects of sensors are considered and the benefits and limitations of each sensor are assessed. Oil spill response often requires that remote sensing be used to detect, map and track the spill of interest. A wide variety of technologies are covered.

► **The Gravity Recovery and Climate Experiment (GRACE) Detection of Ice Mass Loss, Terrestrial Mass Changes, and Ocean Mass Gains** – The gravity field of Earth, caused by the distribution of masses inside and on the surface of Earth, changes in time due to the redistribution of mass. Such mass fluxes can be due both to natural processes (such as the seasonal water cycle, ocean dynamics, or atmospheric variations), as well as due to human actions, such as the systematic withdrawal of groundwater for human consumption.

The GRACE satellite pair has provided the first global measurements at horizontal resolutions from 300 km to global, and time scales from 10 days to interannual. These measurements have been used to assess the mass variability in the oceans, terrestrial water storage, and loss of ice mass in glaciers and ice sheets. Long-term trends, episodic variations such as result from large earthquakes, and seasonal changes have all been measured with unprecedented accuracy and detail.

► **Remote Sensing of Natural Disasters** – Remote sensing involves the use of instruments to study phenomena from a distance. Natural disasters derive from hazards such as volcanoes, flooding, earthquake fires, and weather. Practically speaking, “remote sensing of natural disasters” principally refers to the use of airborne or spaceborne sensors to study natural disasters for detecting, modeling, predicting, analyzing, and mitigating effects on human populations and activities. Satellite instruments can provide imagery of large areas of Earth’s surface and overflights can often easily provide data for hard to access locations (either due to terrain, weather, or causes of the disaster itself). Airborne instruments can also provide much of the same utility from remote sensing with the potential advantages of greater loiter or dwell time and rapid response.

► **Remote Sensing of Ocean Color** – The oceans cover over 70% of Earth’s surface and the life inhabiting the oceans play an important role in shaping Earth’s climate. Phytoplankton, the microscopic organisms in the surface ocean, are responsible for half of the photosynthesis on the planet. These organisms at the base of the food web take up light and carbon dioxide and fix carbon into biological structures releasing oxygen. Estimating the amount of microscopic phytoplankton and their associated primary productivity over the vast expanses of the ocean is extremely challenging from ships. However, as phytoplankton take up light for photosynthesis, they change the color of the surface ocean from blue to green. Such shifts in ocean color can be measured from sensors placed high above the sea on satellites or aircraft and are called “ocean color remote sensing.” In open ocean waters, the ocean color is predominantly driven by the phytoplankton concentration and ocean color remote sensing has been used to estimate the amount of chlorophyll *a*, the primary light-absorbing pigment in all phytoplankton.

► **Earth System Environmental Literacy** – This chapter details the standards of ecological, environmental, and science literacy needed by the public to support decisions involved with assessing climate change data and predictions and taking the required actions.

► **Contemporary Sea Level Variations, Observations and Causes** – Sea-level change is a very sensitive index of climate change and variability. For example, as the ocean warms in response to global warming, seawaters expand, and thus sea level rises. When mountain glaciers melt in response to increasing air temperature, sea level rises because of freshwater mass input to the oceans. Similarly, ice mass loss from the ice sheets causes sea-level rise. Corresponding increases in freshwater transport into the oceans changes water salinity, hence seawater density as well as ocean circulation that in turn affects sea level and its spatial variability. Since the early 1990s, satellite altimetry has become the main tool for precisely and continuously measuring sea level with nearly global coverage and a few days revisit time. Compared to tide gauges which provide sea level relative to the ground, satellite altimetry measures “absolute” sea-level variations. The main factors causing current global mean sea-level rise are thermal expansion of seawaters, land ice loss, and freshwater mass exchange between oceans and land water reservoirs. These contributions vary in response to natural climate variability and to global climate change induced by anthropogenic greenhouse gas emissions.

► **Volcanoes, Observations and Impact** – Volcanoes are critical geologic hazards that challenge our ability to make long-term forecasts of their eruptive behaviors. They also have direct and indirect impacts on human lives and society. As is the case with many geologic phenomena, the time scales over which volcanoes evolve greatly exceed that of a human lifetime. On the other hand, the time scale over which a volcano can move from inactivity to eruption can be rather short: months, weeks, days, and even hours. Thus, the scientific study and monitoring of volcanoes is essential to mitigate risk as well as measure the influx of gasses and particulates into the atmosphere. There are thousands of volcanoes on Earth, and it is impractical to study and implement ground-based monitoring at them all. Fortunately, there are other effective means for volcano monitoring, including increasing capabilities for satellite-based

technologies. The focus of this chapter is mainly on geophysical observations of volcanoes and calderas as they pertain to eruption forecasting and prediction.

Earthquake Faulting, Ground Motions and Deformations

ÖMER AYDAN

Geomechanics and Geoengineering, Tokai University,
Ocean Research Institute, Shizuoka, Japan

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Ground Motions
Ground Deformations
Effects of Surface Ruptures on Structures
Future Directions
Bibliography

Glossary

Damage The damage is defined as the disruption of the integrity/failure of a structure or function of a system.

Earthquake fault It is the fault, along which relative ground displacement occur during an earthquake.

Engineering structures Structures (i.e., dam, tunnel, slope cut, pylon, bridge, viaduct, etc.) with a given function or purpose are those designed and built according to the principles of engineering.

Ground motions Ground motion is a change of a material point in space in a given time and is observed as displacement, velocity or acceleration and they are caused by the application of forces on a given material body (i.e., solid, fluid).

Definition of the Subject and Its Importance

Ground motion characteristics, deformation, and surface breaks of earthquakes depend upon the causative faults. Their effects on the seismic design of engineering structures are not considered in the present codes of

design, although there are attempts to include these factors in some countries (i.e., USA, Japan, Taiwan). This chapter describes ground motions and the effect of surface ruptures associated with earthquake faulting on response and stability of engineering structures.

Introduction

Earthquakes are known to be one of the natural disasters resulting in the huge losses of human lives and properties as experienced in the recent earthquakes. Since there is no way to prevent the occurrence of earthquakes in earthquake-prone countries such as Turkey, Japan, USA, and Taiwan, the design of structures and residential and industrial developments must be done according to possible faulting types of earthquakes (Fig. 1). It is well known that ground motion characteristics, deformation, and surface breaks of earthquakes depend upon the causative faults [1–13]. While many large earthquakes occur along the subduction zones, which are far from the land, and their effects appear as severe shaking, the large inland earthquakes may occur just beneath or nearby urban and industrial zones as observed in the recent great earthquakes.

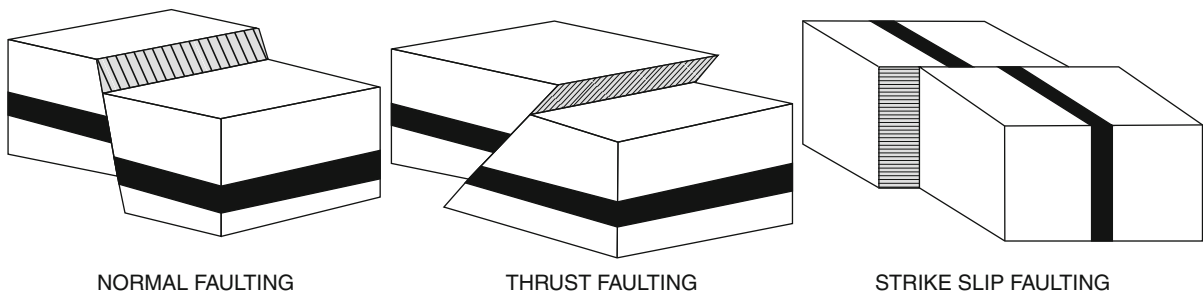
The seismic design of engineering structures is generally formed by considering the possible shaking characteristics of the ground during earthquakes in a given region. It is a fact that the residual (permanent) relative displacement of the ground is not considered in any seismic code all over the world, except for very long linear structures such as pipelines. This problem is currently considered to be beyond the capability of seismic design concept for structures in the earthquake engineering, although it must be dealt with somewhat [1, 2].

In this chapter, the author first describes ground motions and deformations in view of laboratory experiments on rock samples and recordings in earthquakes having different faulting mechanisms. Then the effects of surface ruptures and deformations due to earthquake faulting on the response and stability of engineering structures through observations in recent great earthquakes are presented and discussed.

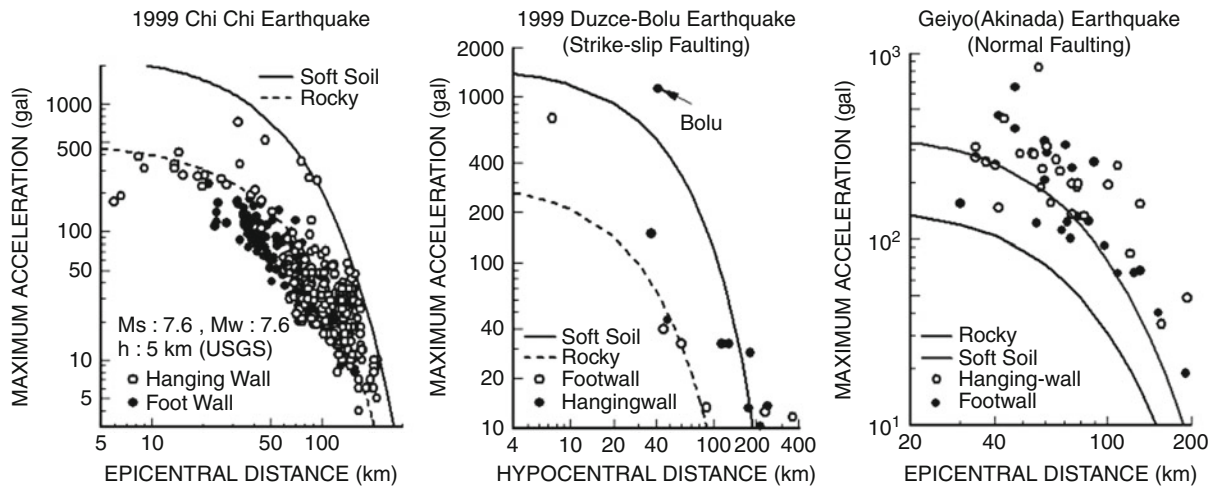
Ground Motions

It is observationally known that the ground motions induced by earthquakes could be much higher in the hanging-wall block or mobile side of the causative fault, as observed in the recent earthquakes such as the 1999 Kocaeli earthquake (strike-slip faulting), the 1999 Chi-chi earthquake (thrust faulting), the 2004 Chuetsu earthquake (blind thrust faulting), and the 2000 Shizuoka earthquake and L'Aquila earthquake (normal faulting) [7, 10, 12–15]. Figure 2 illustrates the effect of hanging-wall effect on the maximum ground accelerations observed in 1999 Chi-chi earthquake (Taiwan), 1999 Düzce earthquake (Turkey), and 2001 Geiyo earthquake (Japan) with different faulting mechanisms [9].

The recent advances in measurement, monitoring, and logging technologies enable measurement and monitoring of the dynamic responses of geomaterials during fracturing and slippage. Studies concerning the dynamic responses of geomaterials during fracturing and slippage can now be more easily undertaken than in the past. Such studies have been recently undertaken at several sites [1, 6–9, 16–19]. The experiments have been performed on geomaterials ranging from very soft materials to hard rocks such as siliceous sandstone by



Earthquake Faulting, Ground Motions and Deformations. Figure 1
Fault types [1]



Earthquake Faulting, Ground Motions and Deformations. Figure 2
Attenuation of maximum ground accelerations for some earthquakes

using different loading schemes and loading frames and specially designed stick-slip experimental device.

Aydan and Ohta [3, 9] reported some tests on samples of various rocks such as Ryukyu limestone, tuff, granite, porphyrite, andesite, and sandstone. Two examples are Fuji-TV No.1 and Mitake Sandstone MS2. Fuji-TV No.1 granite is a prismatic sample ($100 \times 100 \times 200$ mm) (Fig. 3). The acceleration responses start to develop when the applied stress exceeds the peak strength, and it attains the largest value just before the residual state is achieved as seen in Fig. 3. Another important aspect is that the acceleration of the mobile platen is much larger than that of the stationary platen. This is a common feature in all experiments. In another words, the amplitude of accelerations of the mobile part of the loading system is higher than that of the stationary part.

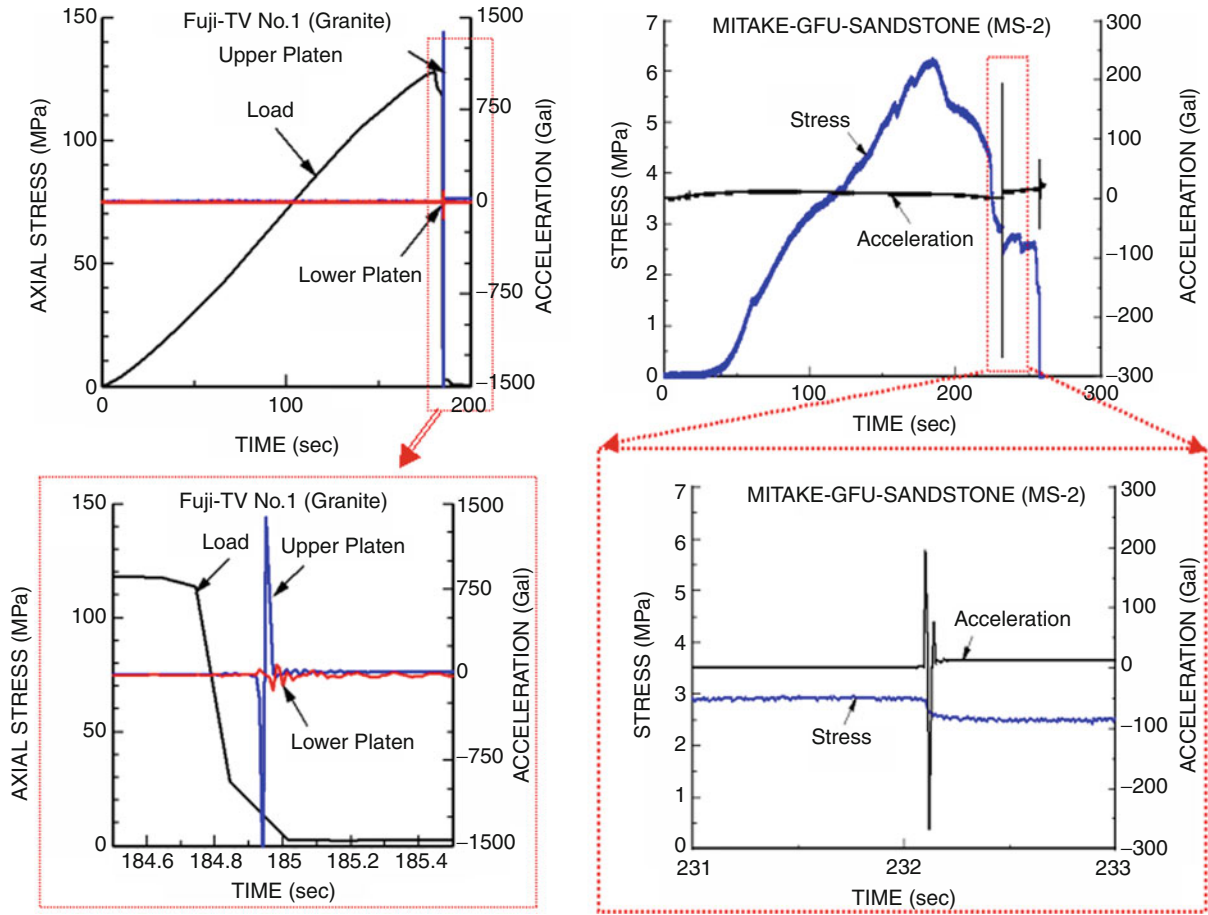
It is well known that the ground motions are generally smaller than those at ground surface. Nasu [20] carried out first instrumental studies on tunnels during the aftershock activity following 1924 Izu earthquake with a 2.4-m offset. Kanai and Tanaka [21] measured ground acceleration in underground caverns and at the ground surface. These measurements indicated that the surface acceleration was generally twice or greater than twice of that at depth as expected theoretically.

Figure 4 shows the records of accelerations at the ground surface and at bedrock 260 m below at

Ichinoseki strong-motion station (IWTH25) of KIK-NET [22] strong-motion network of Japan measured during the 2008 Iwate-Miyagi earthquake. The strong-motion station was located on the hanging-wall side of the fault, and it was very close to the surface rupture. As noted from the figure, the ground acceleration of the UD component was amplified by 5.67 times that at the bedrock. This record is also the highest strong motion recorded in the world so far.

Figure 5 shows the acceleration records measured on the ground surface (GSA) and underground gallery (GSG) during the 2009 Mw 6.3 L'Aquila earthquake [14, 23]. The GSA station is at Assergi, and the GSG station is located in an underground gallery of Gran Sasso Underground Physics Laboratory of Italy. Both stations are founded on Eocene limestone with a shear wave velocity of 1 km/s. Although the epicentral distances and ground conditions are almost the same, the acceleration at ground surface is amplified almost 6.4 times that in the underground gallery.

There are many empirical attenuation relations for estimating ground motions [24–28]. Including the possible next generation attenuation (NGA) relations, all these equations are essentially spherical or cylindrical attenuation relations and cannot take into account directivity effects. As it is shown in the beginning of this section, ground motions such as maximum ground acceleration (AMAX) and maximum ground



Earthquake Faulting, Ground Motions and Deformations. Figure 3
Acceleration and axial response of sandstone and granite samples

velocity (VMAX) have strong directivity effects in relation to fault orientation. Furthermore, these relations are generally far below the maximum ground acceleration, and they are incapable of obtaining the maximum ground acceleration (AMAX) or the preferred term “peak ground acceleration (PGA).”

Aydan and Ohta [4] proposed an attenuation relation by combining their previous proposals [5, 25, 26] together with the consideration of the inclination and length of earthquake fault using the following functional form (Fig. 6):

$$\alpha_{\max} = F_1(V_s) * F_2(R, \theta, \varphi, L^*) * F_3(M) \quad (1)$$

where V_s , θ , φ , L^* , and M are the shear velocity of ground, the angle of the location from the strike and dip of the fault (measured anticlockwise with the

consideration of the mobile side of the fault), and earthquake magnitude. L^* (in km) is a parameter related to the half of the fault length.

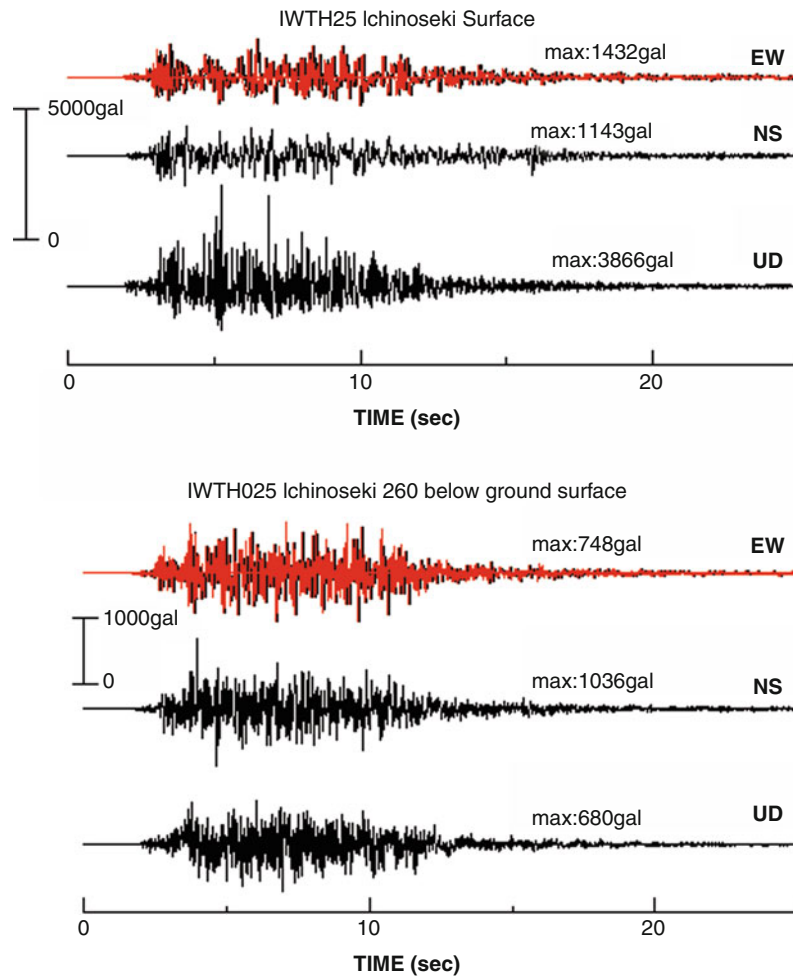
The following specific forms of functions in Eq. 1 were put forward as:

$$F_1(V_s) = Ae^{-V_s/B} \quad (2a)$$

$$F_2(R, \theta, \varphi, L^*) = e^{-R(1-D \sin \theta + E \sin^2 \theta)(1+F \cos \varphi)/L^*} \quad (2b)$$

$$F_3(M) = e^{M/G} - 1 \quad (2c)$$

The same form can be also used for estimating the maximum ground velocity ($V_{\max}$). The constants of the functions for maximum ground acceleration and velocity for intraplate earthquakes without consideration of faulting sense are given in Table 1.



Earthquake Faulting, Ground Motions and Deformations. Figure 4

Acceleration records at ground surface and bedrock at Ichinoseki strong-motion station IWTH25 of KIK-NET in Iwate-Miyagi earthquake

For interplate and slab earthquakes, only the value of constant G is different, and it has the value of 1.05, while other parameters remain the same.

L^* , which is approximately associated with the half of the fault length, is related to the moment magnitude in the following form:

$$L^* = a + be^{cM_w}. \quad (3)$$

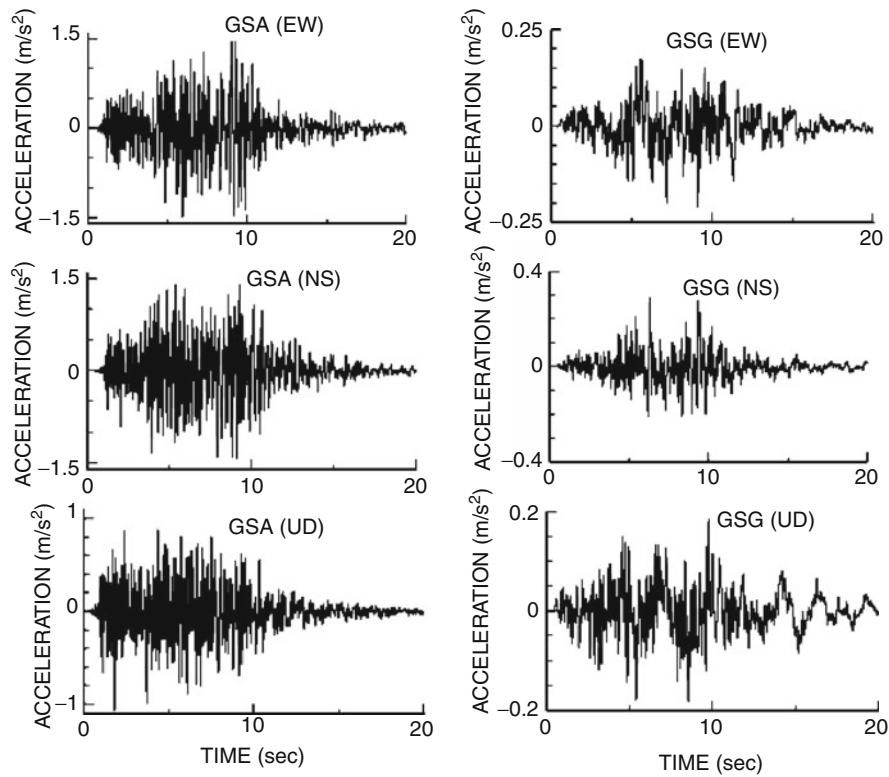
This parameter easily accounts for the effect of fault length. A similar form was initially proposed by Aydan et al. [18] for the effects of earthquake faults on underground structures. However, it should be noted that the attenuation characteristics of the earth's crust vary

depending upon the location and the sense of faulting. The values of coefficients of Eq. 3 are given in Table 2.

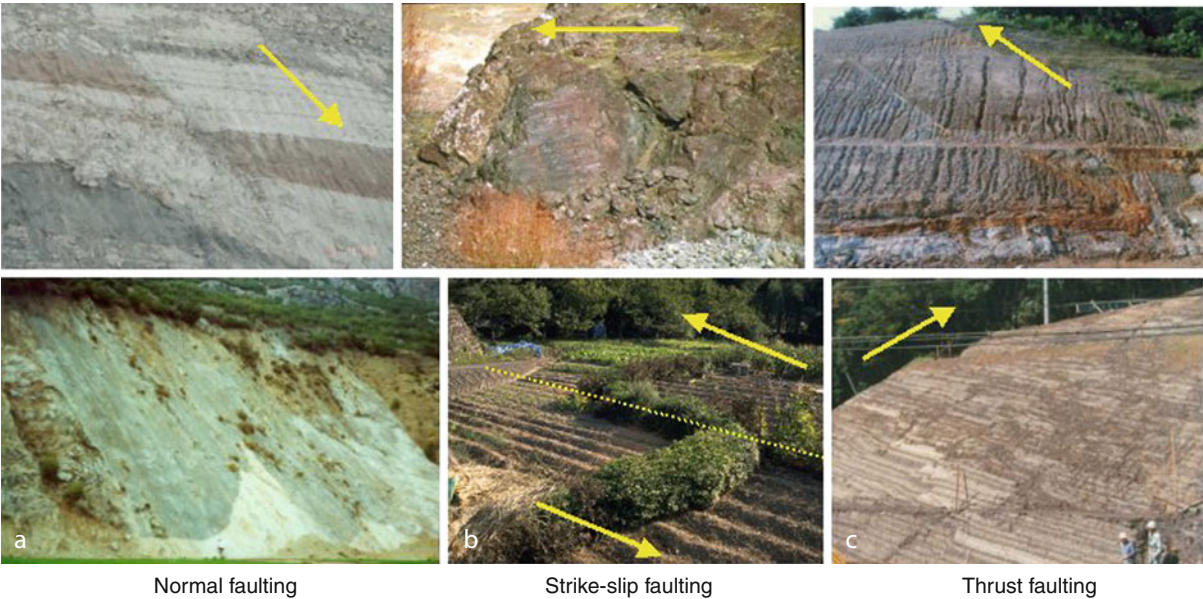
Equation 1 takes into account the location of observation points with respect to earthquake fault and ground conditions, which are fundamentally the main causes of scattering of observational data seen in spherical models for attenuation of ground motions including NGA relations [24–28].

Ground Deformations

A fault is geologically defined as a discontinuity in geological medium along which a relative displacement



Earthquake Faulting, Ground Motions and Deformations. Figure 5
Acceleration records at GSA and GSG strong-motion stations [14]



Earthquake Faulting, Ground Motions and Deformations. Figure 6
Some examples of faulting: (a) normal faulting; (b) strike-slip faulting; (c) thrust faulting

Earthquake Faulting, Ground Motions and Deformations. Table 1 Values of constants in Eq. 2 for intraplate earthquakes

	A	B(m/s)	D	E	F	G(M)	
						(Ms)	(Mw)
AMAX	2.8	1,000	0.5	1.5	0.5	1.11	1.16
VMAX	0.4	1,000	0.5	1.5	0.5	1.11	1.16

took place. Faults are broadly classified into three big groups, namely, normal faults, thrust faults, and strike-slip faults as seen in Figs. 1 and 6. A fault is geologically defined as active if a relative movement took place in a period of less than 2 million years previously. It is well known that a fault zone may involve various kinds of fractures, and it is a zone having a finite volume [29]. The appearance of ground breaks is closely related to geological structure, characteristics of sedimentary deposits, their geometry, the magnitude of earthquakes, and fault movements.

It is well known that the earth's crust is ruptured and contains numerous faults and various kinds of discontinuities, and it is almost impossible to find a piece of land without faults. During the construction of structures such as tunnels, dams, power plants, roadways, railways, power transmission lines, bridges, elevated expressways, etc., it is almost impossible not to cross a fault or faults. Therefore, one of the most important items is how to identify which fault segments observed on ground surface will move or rupture during an earthquake. A fault zone may involve various kinds of fractures as illustrated in Fig. 7, and it is a zone having a finite volume [29]. In other words, a fault is not a single plane. Furthermore, faults may have a negative or positive flower structure as a result of their transtensional or transpressional nature and the reduction of vertical stress near the earth's surface as shown in Fig. 8. For example, even a fault having a narrow thickness at depth may cause a quite broad rupture zone and numerous fractures on the ground surface during earthquakes. Furthermore, the movements of a fault zone may be diluted if a thick alluvial deposit is found on the top of the fault (i.e., 1992 Erzincan earthquake [30]).

The recent global positioning system (GPS) also showed that permanent deformations of the ground

Earthquake Faulting, Ground Motions and Deformations. Table 2 Values of constants in Eq. 3

Faulting type	a	b	c
Normal faulting	30	0.002	1.35
Strike-slip faulting	20	0.002	1.40
Thrust faulting	30	0.003	1.45

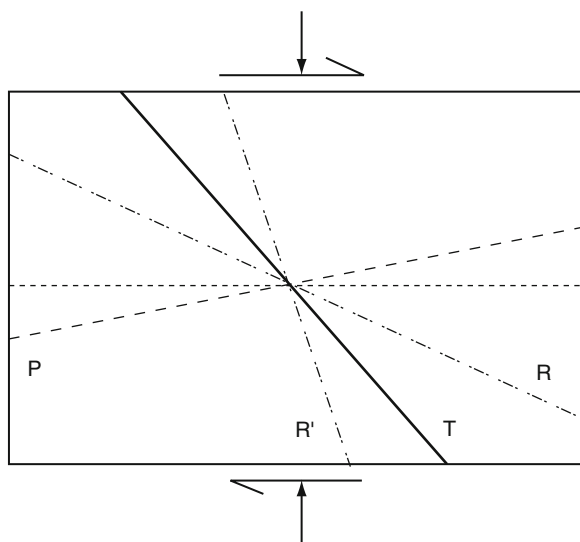
surface occur after each earthquake (Fig. 2). The permanent ground deformation may result from different causes such as faulting, slope failure, liquefaction, and plastic deformation induced by ground shaking [18]. These types of ground deformations will have limited effect on small structures as long as the surface breaks do not pass beneath those structures. However, such deformations may cause tremendous forces on long and/or large structures. The ground deformation may induce large tensile or compression forces, as well as bending stresses, in structures depending upon the character of permanent ground deformations. As an example, the ground deformations reported by Reilinger et al. [31] are shown in Fig. 9, which were caused by a strike-slip fault during the 1999 Kocaeli earthquake in Turkey. Blind faults and folding processes may also induce some peculiar ground deformations and associated folding of soft overlaying sedimentary layers. Such deformations caused tremendous damage on tunnels during the 2004 Chuetsu earthquake, although no distinct rupturing took place.

Effects of Surface Ruptures on Structures

In this section, typical examples of damage to various structures induced by the fault breaks observed in recent large earthquakes since 1995 are presented [1, 2, 11, 32–50].

Roadways and Railways

The Trans-European Motorway (TEM) was damaged at three different locations by the earthquake fault caused by the 1999 Kocaeli earthquake. The motorway with east and west bounds having three lanes each was slightly elevated through embankments in the earthquake-affected region. The surfacing of the motorway was damaged by rupturing and buckling (Fig. 10). The



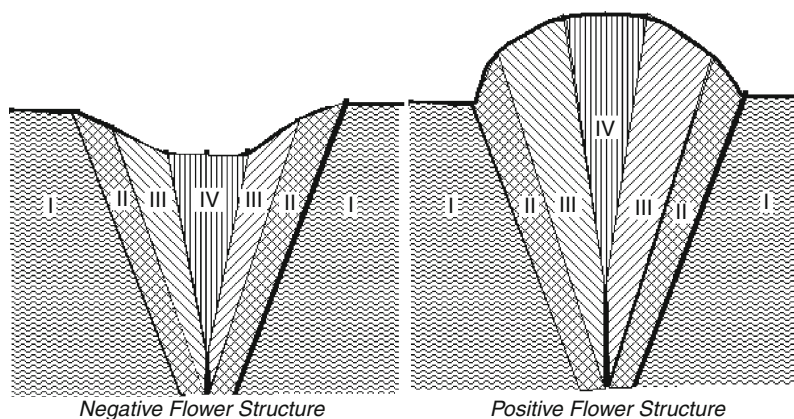
Earthquake Faulting, Ground Motions and Deformations. Figure 7

Fractures in a shear zone or fault [29]

railways were originally built on existing ground surface but were buckled near Tepetarla station where the earthquake fault crossed the railways at an angle of $50\text{--}55^\circ$ with well-known “S” shape (Fig. 11).

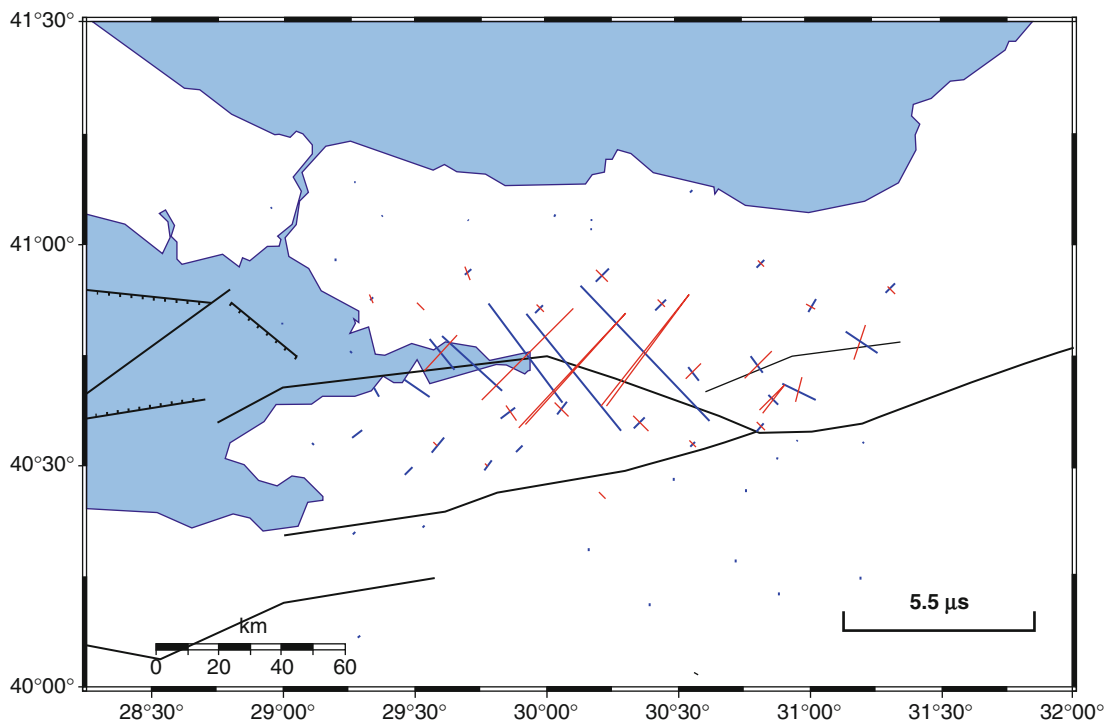
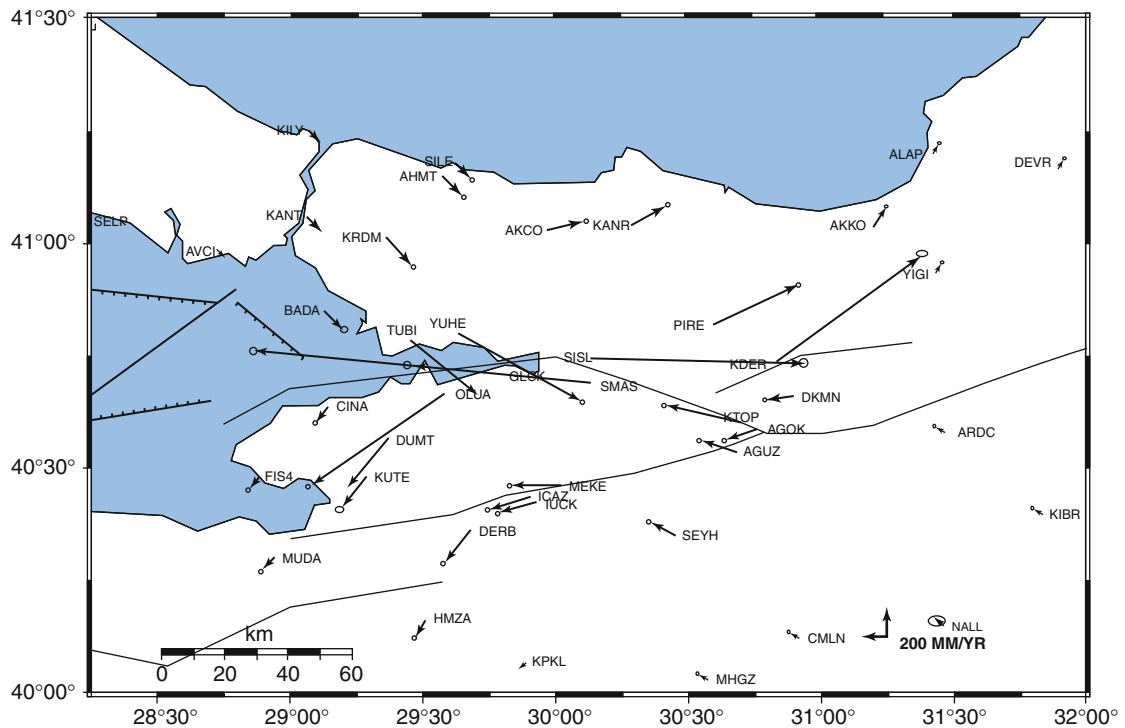
Bridges and Viaducts

Along the damaged section of the TEM motorway mentioned above, there were several overpass bridges. Among them, a four-span overpass bridge at Arifiye junction collapsed as a result of faulting (Fig. 12). The fault rupture passed between the northern abutment and the adjacent pier. The overpass was designed as a simply supported structure according to the modified AASHTO standards, and girders had elastomeric bearings. However, the girders were connected to each other through prestressed cables. The angle between the motorway and the strike of the earthquake fault was approximately 15° , while the angle between the axis of the overpass bridge and the strike of the fault was 65° . The measurements of the relative displacement in the vicinity of the fault range between 330 and 450 cm. Therefore, an average value of 390 cm could be assumed for the relative displacement between the pier and the abutment of the bridge. The Pefong bridge collapsed due to thrust faulting in the 1999 Chi-chi earthquake, which passed between the piers near its southern abutment as seen in Fig. 13.



Earthquake Faulting, Ground Motions and Deformations. Figure 8

Negative and positive flower structures due to transtension or transpression faulting (Modified from [2, 11, 29])



Earthquake Faulting, Ground Motions and Deformations. Figure 9

Permanent ground deformations and associated straining induced by the 1999 Kocaeli earthquake



Earthquake Faulting, Ground Motions and Deformations. Figure 10

Buckling of roadway surfacing



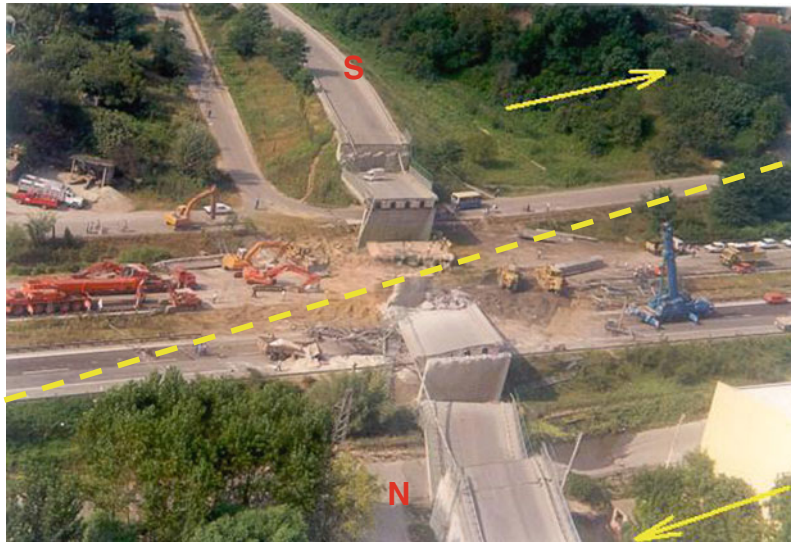
Earthquake Faulting, Ground Motions and Deformations. Figure 11

Buckled railways

Dams

The Shihkang dam, which is a concrete gravity dam with a height of 25 m, was ruptured by thrust-type faulting during the 1999 Chi-chi earthquake (Fig. 14). The relative displacement between the uplifted part of the dam was more than 980 cm. Liyutan rock-fill dam with a height of 90 m and a crest width of 210 m, which

was on the overhanging block of Chelungpu fault, was not damaged even when the acceleration records at this dam showed that the acceleration was amplified 4.5 times of that at the base of the dam (105 Gal). The deformation zone of faulting during the 2008 Wenchuan earthquake caused some damage at Zipingpu dam with concrete facing.



Earthquake Faulting, Ground Motions and Deformations. Figure 12

The collapse of the overpass bridge at Arifiye



Earthquake Faulting, Ground Motions and Deformations. Figure 13

Collapse of Pefong bridge (note the uplifted ground on RHS)

Tunnels

The past experience on the performance of tunnels through active fault zone during earthquakes indicates that the damage is restricted into certain locations [18, 41–43]. Portals and the locations where the tunnel

crosses the fault may be damaged as had occurred in the 2004 Chuetsu, 2005 Kashmir, and 2008 Wenchuan earthquakes (Fig. 15). A section nearby the Elmalık portal of Bolu tunnel collapsed (Fig. 16). This section of the tunnel was excavated under very heavy squeezing conditions [51]. Well-known examples of damage to



Earthquake Faulting, Ground Motions and Deformations. Figure 14
Failure of Shihkang dam due to thrust faulting



Earthquake Faulting, Ground Motions and Deformations. Figure 15
Examples of damaged portals of tunnels



Earthquake Faulting, Ground Motions and Deformations. Figure 16
Collapse of Bolu tunnel during the 1999 Düzce earthquake

tunnels at locations, where the fault rupture crossed the tunnel, are mainly observed in Japan. The Tanna fault ruptured during the 1930 Kita-Izu earthquake and caused damage to a railway tunnel, with a relative displacement of about 100 cm [52, 53]. The 1978 Izu-Oshima Kinkai earthquake induced damage to Inatori railway tunnel [54, 55]. A similar type of damage with a small amount of relative displacements due to motions of Rokko, Egeyama, and Koyo faults to the tunnels of Shinkansen and subway lines through Rokko mountains was also observed [33]. During the 1999 Chi-chi earthquake, the portal of the water intake tunnels was ruptured for a distance of 10 m as a result of thrust faulting. Except this section, the tunnel was undamaged for its entire length.

Jiujiaya tunnel is a 2,282-m-long double-lane tunnel 226.6 km away from the earthquake epicenter and about 3–5 km away from the earthquake fault of Wenchuan earthquake. The tunnel face was 983 m from the south portal at the time of the earthquake. The concrete lining follows the tunnel face at a distance of approximately 30 m. Thirty workers were working at the tunnel face,

and one worker was killed by the flying pieces of rock bolts, shotcrete, and bearing plates caused by intense deformation of the tunnel face during the earthquake [17, 18]. The concrete lining was ruptured and fell down at several sections (Fig. 17). The effect of the unreinforced lining rupturing was quite large and intense in the vicinity of the tunnel face. The rupturing of the concrete lining generally occurred at the crown sections, although there was rupturing along the shoulders of the tunnel at several places. Furthermore, the invert was uplifted due to buckling at the middle sections.

Subways

The behavior of subways in active fault zones is basically quite similar to that of tunnels. The Daikai station of the subway line in Kobe was caused by the lateral strike-slip movement of Egeyama fault just beneath this station, although some have tried to associate the collapse of the station with the intensity of shaking (Fig. 18). The investigation of the collapse of this station showed that collapse was not due to shaking, as the

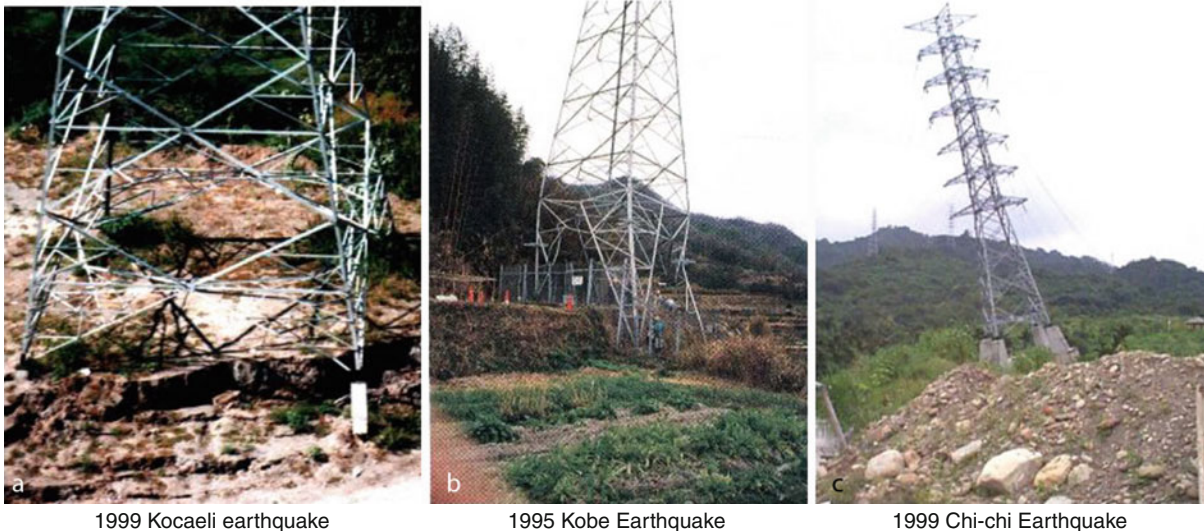


Earthquake Faulting, Ground Motions and Deformations. Figure 17

Earthquake damage at Jiujiaya tunnel due to permanent deformations [18]



Earthquake Faulting, Ground Motions and Deformations. Figure 18
The collapse of Daikai station



Earthquake Faulting, Ground Motions and Deformations. Figure 19
Damage to pylons due to faulting. (a) 1999 Kocaeli earthquake. (b) 1995 Kobe earthquake. (c) 1999 Chi-chi earthquake

central columns of the station were subjected to torsional failure due to the permanent ground displacement, consistent with the lateral strike-slip movement of Egeyama fault [33].

Power Transmission Lines

Power transmission lines generally consist of pylons and power transmission cables. The design of pylons

and cables is generally based on the wind loads resulting from typhoons or hurricanes. The cables do not fail during earthquakes unless the pylons are toppled due to faulting, shaking, or slope failure. During the 1999 Kocaeli earthquake, only one pylon was damaged, at a site near Ford-Otosan automobile factory at Kavaklı district of Gölcük town. At this site, a normal fault, secondary to the main lateral strike-slip faulting event, crossed through the foundations of the pylon

and with a vertical lift of about 240 cm. One of the foundations of the pylon was pulled out of the ground and exposed (Fig. 19a). Some of its truss elements were slightly buckled. A similar type of damage to pylons straddling the Nojima fault break in the 1995 Kobe earthquake (Fig. 19b) and the Chelungpu fault during the 1999 Chi-chi earthquake has been observed as shown in Fig. 19c.

Line-Like and Tubular Structures

Tubular structures may be specifically designated as petrol and gas pipelines, water pipes, and sewage systems. They can be also classified as line-like structures.

These structures may fail either by buckling or separation during a faulting event. Five such incidents were observed during the 1999 Kocaeli earthquake (Fig. 20). One of the incidents involved separation of a ductile iron pipe as a result of faulting near the collapsed overpass bridge. The second incident took place at the pumping facility of the Seka paper mill plant at Sapanca Lake. The third incident occurred near Tepetarla village where the railways buckled. The fourth and fifth incidents took place at Arifiye and nearby Başiskele. The fifth incident was quite important, since the fault caused a heavy damage to the main water pipe having a diameter of 2 m. Similar types of failures took place in the sewage pipe



Earthquake Faulting, Ground Motions and Deformations. Figure 20
Damage to pipes in the 1999 Kocaeli and 1999 Düzce earthquakes

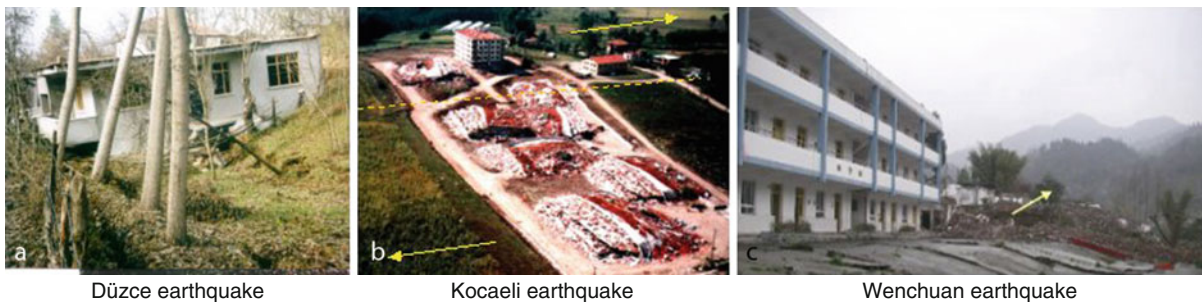
networks whenever faulting breaks were observed. The natural gas pipelines crossing the İzmit gulf between Yalova and Pendik were undamaged. Some brittle asbestos water pipes were also damaged in Kaynaşlı and Fındıklı due the fault rupture of the 1999 Düzce earthquake.

Buildings

Many buildings along the earthquake faults of the 1999 Kocaeli, Düzce, Chi-chi, and Wenchuan earthquakes were affected in a similar manner. During the site investigation of the 1999 Kocaeli earthquake, one could see totally collapsed, severely damaged, or intact buildings on or adjacent to the traces of the fault breaks. Examples are many; it is difficult to note all of them. Two typical examples are briefly discussed. Nevertheless, this topic deserves more thorough investigation. The first example is a single-story reinforced concrete house with a raft foundation in Fındıklı village. The fault passed just underneath the building (Fig. 21a). The relative displacement of the fault break was about 200 cm with a 100-cm downthrow. The building was tilted, but no damage was observed. A very peculiar type of damage at an apartment complex consisting of 8 five-story apartment blocks was observed in Kullar village as shown in Fig. 21b. Seven apartment blocks failed in a pancake mode, while one apartment block remained standing. One of the failed apartment blocks just crossed the fault break which has a relative horizontal displacement of 240 cm and 20–25 cm vertical throw (north side down).

The ground surface sloped to north. One of two apartment blocks on the southern side was damaged, while the other one collapsed toward the east in a pancake mode in accordance with the movement of its foundation. The five blocks on the northern side completely collapsed in a pancake mode toward west in accordance with the direction of shaking. Except for the apartment block over the fault break, the failure of the five apartment blocks on the northern side of the fault break may be considered to be purely due to shaking. The Bailu secondary school building of reinforced concrete survived the Wenchuan earthquake, even though the surface rupture of the thrust fault passed underneath the building (Fig. 21c). Although the intensity of shaking on the southern side of the fault break should be the same, the damaged self-standing apartment block deserves further consideration. Whatever the reason, it is of great interest that even the most vulnerable buildings may also survive within a distance of 5–6 m from the fault break during inland earthquakes.

Landslides and Rockfalls The recent large earthquakes caused megascale slope failures and rockfalls, particularly along the surface ruptures on the hanging-wall side of faults. The slope failure induced in Beichuan town during the 2008 Wenchuan earthquake [40, 50] is of great interest. In association with the sliding motion of the earthquake fault, NW- or SE-facing slopes failed during this earthquake. There were two large-scale slope failures (landslides) in Beichuan town, which destroyed numerous buildings



Earthquake Faulting, Ground Motions and Deformations. Figure 21

The behavior of reinforced concrete buildings over surface ruptures. (a) Düzce earthquake. (b) Kocaeli earthquake. (c) Wenchuan earthquake



Earthquake Faulting, Ground Motions and Deformations. Figure 22
Views of landslides in Beichuan

and facilities. The NW-facing landslide (Jingjiashan) involved mainly limestone, while the SE-facing landslide (Wangjiaya) involved phyllite (mudstone, according to some) rock unit (Fig. 22). Limestone layers dip toward the valley side with an inclination of about 30° . There are several faults dipping parallel to the failure surface within the rock mass. The angles of the lower and upper parts of the failed slope are 60° and $30\text{--}35^\circ$, respectively. The existence of several faults dipping parallel to the slope with an inclination of about $60\text{--}65^\circ$ creates a stepped failure surface.

The SE-facing slope (Wangjiaya landslide) may involve slippage along the steeply dipping bedding plane (fault plane?) and shearing through the layered rock mass. In other words, it may be classified as a combined sliding and shearing effect [56]. The angles of the lower and upper parts of the failed slope are $40\text{--}45^\circ$ and $30\text{--}35^\circ$, respectively. The layers dip at an angle of 40° toward the valley, and the shearing plane is inclined at an angle of 20° .

Future Directions

The characteristics of ground motions and deformations associated with earthquake faulting are presented in this chapter. Furthermore, the response and damage to structures due to surface ruptures are illustrated through actual observations in recent great earthquakes.

The advanced strong-motion attenuation relations by Aydan and Ohta [4] seem to explain the wide range of observational data, which are considered to be scattering when spherical or cylindrical symmetric attenuation relations are used. A multisegmented or multisource model may be used for estimating strong motions [4] if surface ruptures due to earthquakes are relatively long.

The GPS measurements of ground deformations during earthquakes ($M > 6$) clearly indicated that permanent ground deformations do occur. The overall deformation response involves both coseismic and

time-dependent components. The permanent ground deformation may result from different causes such as faulting, slope failure, liquefaction, and plastic deformation induced by ground shaking. They may cause tremendous forces on long and/or large underground structures such as tunnels, powerhouses, and underground storage facilities of oil, gas, and nuclear wastes. Therefore, the current seismic design codes must be revised to include these effects.

As a conclusion, *it is almost impossible for mankind to prevent the damage to structures if a fault break happens to be just passing underneath the structures.* Nevertheless, the adverse effects of fault breaks on structures can be minimized to a negligible level in view of findings from the actual examples presented in this chapter. The following suggestions may be put forward for developing the seismic codes for structures in active fault zones:

1. Roadways and railways should not be elevated, and they should be constructed on the existing ground surface in active fault zones. In doing so, recovery of roadways and railways damaged by fault breaks in the event of the earthquakes would be very rapid. As a result, effective rescue and rehabilitation works would be possible.
2. Bridges and viaducts should be constructed as redundant structures, and simply supported structures should be avoided. Although prestressing can be used to increase the confinement of girders along the longitudinal axis of the bridge, the structures should not be wholly prestressed to prevent domino action in the event of failure. Furthermore, T-type piers must not be used as they constitute top-heavy situations which are undesirable in case of faulting as well as shaking caused by earthquakes.
3. Dams should never be built on active faults defined in engineering sense because of the risk of severe impacts to structures and residential areas downstream in the event of failure. If for some reason dams must be constructed in faulted zones, their height should be kept to a minimum and the type of dams should be rock-fill.
4. Observations and studies on various underground openings show that they are resistant to shaking. Nevertheless, the existence of discontinuities makes them vulnerable to collapse, particularly in case of shallow underground openings. This may have some important implications on areas where shallow abandoned mines, underground shelters, and old tunnels exist. As damage to tunnel linings due to fault breaks would be localized, there is no need for extra precautions at those locations. Underground openings crossing faults and fracture zones may be enlarged to accommodate relative slips along faults and fracture zones. The lining of the openings should be ductile to accommodate permanent ground deformations at such zones. Furthermore, the brittle linings of the existing underground structures should be lined with ductile thin plates or fiber-reinforced polymers together with rock bolts at fracture and fault zones, where permanent ground deformations may occur.
5. Tubular structures should be designed with the use of flexible joints to accommodate relative displacements with the consideration of likely relative displacement of the ground.
6. Buildings should be designed as rigid box-like structures with the use of shear walls. Walls with fragile hollow bricks having no structural strength must not be allowed during building construction. If the brick walls are to be used, the walls should be constructed before the construction of columns and beams in order to attain the necessary structural integrity, and they should be light, resistant, and ductile.
7. Vertical accelerations of inland earthquakes can be 0.65–0.75 times the maximum horizontal accelerations; sometimes they may even exceed maximum horizontal acceleration. The effect of vertical accelerations must be considered in structural design.
8. Structures of great importance should be sited on the rock foundations rather than on loose alluvial ground, since amplification of movement in loose ground may be several times that of rock foundation. The failure risk of structures during earthquakes could be drastically reduced by such a selection. Nuclear power plants and even arch dams built on rock foundation performed very well without any damage during the recent great earthquakes.
9. Global, regional, and local geological and geotechnical maps of cities and towns for residential and industrial developments in seismically active fault zones should be prepared. Furthermore, detailed

geological and geotechnical investigations should be required before any new development to reduce the seismic risks to structures during earthquakes in active fault zones.

Bibliography

Primary Literature

1. Aydan Ö (2003) Actual observations and numerical simulations of surface fault ruptures and their effects engineering structures. In: The 8th U.S.-Japan workshop on earthquake resistant design of lifeline facilities and countermeasures against liquefaction. Technical report, MCEER-03-0003, pp 227–237
2. Aydan Ö, Ulusay R, Hasgür Z, Hamada M (1999) The behaviour of structures built in active fault zones in view of actual examples from the 1999 Kocaeli and Chi-chi Earthquakes. In: ITU-IAHS international conference on the Kocaeli earthquake, 17 Aug 1999: a scientific assessment and recommendations for re-building, Istanbul, pp 131–142
3. Aydan Ö, Ohta Y (2006) The characteristics of ground motions in the neighbourhood of earthquake faults and their evaluation. In: Symposium on the records and issues of recent great earthquakes in Japan and overseas, EEC-JSCE, Tokyo, pp 114–120
4. Aydan Ö, Ohta Y (2011) A new proposal for strong ground motion estimations with the consideration of characteristics of earthquake fault. In: Seventh national conference on earthquake engineering, Istanbul, paper no 66
5. Aydan Ö (2007) Inference of seismic characteristics of possible earthquakes and liquefaction and landslide risks from active faults (in Turkish). In: The 6th national conference on earthquake engineering of Turkey, Istanbul, vol 1, pp 563–574
6. Ohta Y, Aydan Ö (2004) An experimental study on ground motions and permanent deformation nearby faults. *J Sch Mar Sci Technol* 2(3):1–12
7. Ohta Y (2011) A fundamental research on the effects of ground motions and permanent ground deformations neighborhood earthquake faults on civil engineering structures (in Japanese). Doctorate Thesis, Graduate School of Science and Technology, Tokai University, 272 p
8. Ohta Y, Aydan Ö (2009) An experimental and theoretical study on stick-slip phenomenon with some considerations from scientific and engineering viewpoints of earthquakes. *J Sch Mar Sci Technol* 8(3):53–67
9. Ohta Y, Aydan Ö (2010) The dynamic responses of geomaterials during fracturing and slippage. *Rock Mech Rock Eng* 43(6):727–740
10. Chang T-Y, Cotton F, Tsai Y-B, Angelier J (2004) Quantification of hanging-wall effects on ground motion: some insights from the 1999 Chi-Chi earthquake. *Bull Seismol Soc Am* 94(6):2186–2197
11. Ulusay R, Aydan Ö, Hamada M (2002) The behavior of structures built on active fault zones: examples from the recent earthquakes of Turkey. *Struct Eng/Earthquake Eng JSCE* 19(2):149–167
12. Somerville PG, Smith NF, Graves RW, Abrahamson NA (1997) Modification of empirical strong ground motion attenuation relations to include the amplitude and duration effects of rupture directivity. *Seismol Res Lett* 68: 199–222
13. Tsai YB, Huang MW (2000) Strong ground motion characteristics of the Chi-Chi Taiwan earthquake of 21 Sept 1999. In: 2000 NCHU-Waseda joint seminar on earthquake engineering, 17–18 Jul 2000, Taichung, pp 1–32
14. Aydan Ö, Kumsar H, Toprak S, Barla G (2009) Characteristics of 2009 L'Aquila earthquake with an emphasis on earthquake prediction and geotechnical damage. *J Mar Sci Technol Tokai Univ* 9(3):23–51
15. Abrahamson NA, Somerville PG (1996) Effects of the hanging wall and footwall on ground motions recorded during the Northridge earthquake. *Bull Seismol Soc Am* 86(1B): 593–599
16. Aydan Ö (2003) An experimental study on the dynamic responses of geomaterials during fracturing. *J Sch Mar Sci Technol Tokai Univ* 1(2):1–7
17. Aydan Ö, Ohta Y, Geniş M, Tokashiki N, Ohkubo K (2010) Response and earthquake induced damage of underground structures in rock mass. *J Rock Mech Tunnell Technol* 16(1):19–45
18. Aydan Ö, Ohta Y, Geniş M, Tokashiki N, Ohkubo K (2010) Response and stability of underground structures in rock mass during earthquakes. *Rock Mech Rock Eng* 43(6):857–875
19. Aydan Ö, Daido M, Tokashiki N, Bilgin A, Kawamoto T (2007) Acceleration response of rocks during fracturing and its implications in earthquake engineering. In: 11th ISRM Congress, Lisbon, vol 2, pp 1095–1100
20. Nasu N (1931) Comparative studies of earthquake motions above ground and in a tunnel. *Bull Earthquake Res Inst Tokyo Univ* 9:454–472
21. Kanai K, Tanaka T (1951) Observations of earthquake motion at different depths of the earth. *Bull Earthquake Res Inst Tokyo Univer* 28:107–113
22. K-NET, KiK-Net (2007, 2008) Digital acceleration records of earthquakes since 1998. <http://www.k-net.bosai.go.jp/> and <http://www.kik.bosai.go.jp/>
23. Kawashima K, Aydan Ö, Aoki T, Kishimoto I, Konagai K, Matsui T, Sakuta J, Takahashi N, Teodori SP, Yashima A (2010) Reconnaissance investigation on the damage of the 2009 L'Aquila, Central Italy earthquake. *J Earthquake Eng* 14: 817–841
24. Joyner WB, Boore DM (1981) Peak horizontal acceleration and velocity from strong motion records from the 1979 Imperial Valley California earthquake. *Bull Seismol Soc Am* 71(6): 2011–2038
25. Aydan Ö, Sezaki M, Yazar R (1996) The seismic characteristics of Turkish earthquakes. In: The 11th World Conf. on Earthquake Eng., CD-2, Paper No:1270
26. Aydan Ö (2001) Comparison of suitability of submerged tunnel and shield tunnel for subsea passage of Bosphorus (in Turkish). *Geol Eng J* 25(1):1–17

27. Campbell KW (1981) Near source attenuation of peak horizontal acceleration. *Bull Seismol Soc Am* 71(6):2039–2070
28. Ambraseys NN (1988) Engineering seismology. *Earthquake Eng Struct Dynam* 17:1–105
29. Aydan Ö, Shimizu Y, Akagi T, Kawamoto T (1997) Development of fracture zones in rock masses. In: *Int Symp. on deformation and progressive failure in geomechanics, IS-NAGOYA, Nagoya*, pp 533–538
30. Hamada M, Aydan Ö (1992) A report on the site investigation of the March 13 earthquake of Erzincan, Turkey. ADEP, Association for Development of Earthquake Prediction, 86 p
31. Reilinger RE, Ergintav S, Burgmann R, McClusky S, Lenk O, Barka A, Gürkan O, Hearn L, Feigl KL, Çakmak R, Aktug B, Özener H, Toksöz MN (2000) Coseismic and postseismic fault slip for the 17 Aug 1999, $M = 7.5$, Izmit, Turkey earthquake. *Science* 289:1519–1524
32. Asakura T, Sato Y (1998) Mountain tunnels damage in the 1995 in Hyogo-ken Nanbu earthquake. *Railway Tech Res Inst (RTRI)* 39(1):9–16
33. Aydan Ö (1996) Faulting and characteristics of earthquake waves in Hyogo-ken Nanbu earthquake of 17 Jan 1995 (in Turkish). *Jeoloji Mühendisliği* 48:63–77
34. Aydan Ö, Kawamoto T (2004) The damage to abandoned lignite mines caused by the 2003 Miyagi-Hokubu earthquake and some considerations on its causes. In: *3rd Asian rock mechanics symposium, Kyoto*, pp 525–530
35. Aydan Ö, Kumsar H (2010) An experimental and theoretical approach on the modeling of sliding response of rock wedges under dynamic loading. *Rock Mech Rock Eng* 43(6):821–830
36. Aydan Ö, Hamada M (2006) Damage to civil engineering structures by 8 Oct 2005 Kashmir earthquake and recommendations for recovery and reconstruction. *J Disaster Res* 1(3):1–9
37. Aydan Ö, Hamada M, Suzuki Y (2005) Some observations and considerations on the damage induced by the tsunami of the 2004 Sumatra earthquake on structures and coast. *J Sch Mar Sci Technol* 3(1):79–94
38. Aydan Ö, Miwa S, Kodama H, Suzuki T (2005) The characteristics of M8.7 Nias earthquake of 28 Mar 2005 and induced tsunami and structural damages. *J Sch Mar Sci Technol Tokai Univ* 3(2):66–83
39. Aydan Ö, Ohta Y, Hamada M (2009) Geotechnical evaluation of slope and ground failures during the 8 Oct 2005 Muzaffarabad earthquake in Pakistan. *J Seismol* 13(3):399–413
40. Aydan Ö, Hamada M, Itoh J, Ohkubo K (2009) Damage to civil engineering structures with an emphasis on rock slope failures and tunnel damage induced by the 2008 Wenchuan earthquake. *J Disaster Res* 4(2):153–164
41. Hashimoto S, Miwa K, Ohashi M, Fuse K (1999) Surface soil deformation and tunnel deformation caused by the 3 Sept 1998, Mid-North Iwate earthquake. In: *7th Tohoku regional convention, Japan Society of Engineering Geology*
42. Yashiro K, Kojima Y, Shimizu M (2007) Historical earthquake damage to tunnels in Japan and case studies of railways tunnels in the 2004 Niigata-ken Chuetsu earthquake. *Q RTRI* 48(3):136–141
43. Asakura T, Shiba Y, Sato Y, Iwatate T (1996) Mountain tunnels performance in 1995 Hyogo-ken Nanbu earthquake. Special report of the 1995 Hyogo-ken Nanbu earthquake, Committee of Earthquake Engineering, JSCE
44. Aydan Ö, Kumsar H (1997) A site investigation of 1 Oct 1995 Dinar earthquake. *Turkish Earthquake Foundation, TDV/DR 97–003, Istanbul*, p 115
45. Aydan Ö, Ulusay R, Kumsar H, Sönmez H, Tuncay E (1998) A site investigation of Adana-Ceyhan earthquake of 27 June 1998. *Turkish Earthquake Foundation, TDV/DR 006–30, 131 p*
46. Aydan Ö, Ulusay R, Hasgür Z, Taşkın B (1999) A site investigation of Kocaeli earthquake of 17 Aug 1999. *Turkish Earthquake Foundation, TDV/DR 08–49, 180 p*
47. Aydan Ö, Ulusay R, Kumsar H, Tuncay E (2000a) Site investigation and engineering evaluation of the Düzce-Bolu earthquake of 12 Nov 1999. *Turkish Earthquake Foundation, TDV/DR 095–51, 307 p*
48. Ulusay R, Aydan Ö, Erken E, Kumsar H, Tuncay E, Kaya Z (2003) Site investigation and engineering evaluation of the Cay-Eber earthquake of 3 Feb 2002. *Turkish Earthquake Foundation, TDV/DR 012–79, 213 p*
49. Wang WL, Wang TT, Su JJ, Lin CH, Seng CR, Huang TH (2001) Assessment of damage in mountain tunnels due to the Taiwan Chi-Chi earthquake. *Tunnel Undergr Space Technol* 16:133–150
50. Aydan Ö, Ohta Y, Hamada M, Ito J, Ohkubo K (2009) The response and damage of structures along the fault rupture traces of the 2008 Wenchuan earthquake. In: *International conference on earthquake engineering: the 1st anniversary of Wenchuan earthquake, Chengdu*, pp 625–633
51. Aydan Ö, Dalgıç S, Kawamoto T (2000) Prediction of squeezing potential of rocks in tunnelling through a combination of an analytical method and rock mass classifications. *Italian Geotech J* 34(1):41–45
52. Kuno H (1935) The geologic section along the Tanna tunnel. *Bull Earthquake Res Inst Univ Tokyo* 14:92–103
53. Sakurai T (1999) A report on the earthquake fault appearing in the Tanna tunnel under construction by North-Izu earthquake 1930 (in Japanese). *J Jpn Soc Eng Geol* 39(6):540–544
54. Kawakami H (1984) Evaluation of deformation of tunnel structure due to Izu-Oshima Kinkai earthquake of 1978. *Earthquake Eng Struct Dynam* 12(3):369–383
55. Tsuneishi Y, Ito T, Kano K (1978) Surface faulting associated with the 1978 Izu-Oshima- Kinkai earthquake. *Bull Earthquake Res Inst Univ Tokyo* 53:649–674
56. Aydan, Shimizu Y, Kawamoto T (1992) The stability of rock slopes against combined shearing and sliding failures and their stabilisation. *Int. Symp. on Rock Slopes, New Delhi*, pp 203–210

Books and Reviews

- Fowler CMR (1990) *The solid earth – an introduction to global geophysics*. Cambridge University Press, Cambridge

Okamoto S (1973) Introduction to earthquake engineering. Univ of Tokyo Press, Tokyo

Yeats RS, Sieh K, Allen CR (1997) The geology of earthquakes. Oxford University Press, Oxford/New York

Ecological and Health Risks at Low Doses

KRISTINE L. WILLETT¹, CHRISTY M. FORAN²

¹Department of Pharmacology and Environmental Toxicology Research Program, University of Mississippi, School of Pharmacy, University, MS, USA

²Engineer Research and Development Center-Environmental Lab, Concord, MA, USA

Article Outline

Glossary
Definition of Subject
Introduction
Physiological Response to Low Doses
Specific Examples
Future Directions
Bibliography

Glossary

Acclimation A decline in response with previous exposure.

Adaptation Physiological changes that increase the fitness of a population in response to chronic exposure.

Epigenetic Mechanisms of altered gene expression that do not change primary DNA sequence.

Hormesis A beneficial effect from a low dose of a “poison” or contaminant.

NOAEL No-observed-adverse-effect level is the highest treatment level from which the response does not differ statistically from control groups.

Threshold dose Dose required to produce a measurable response.

Definition of Subject

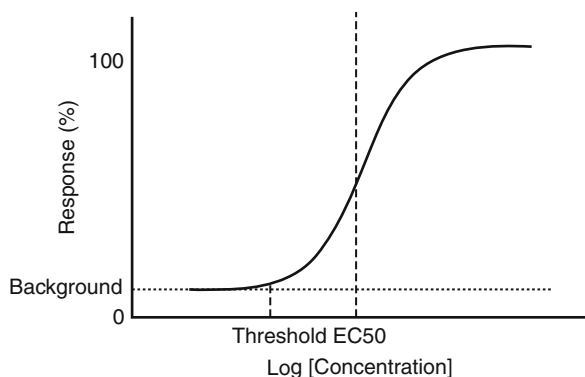
The concept of “dose makes the poison” has been attributed to Paracelsus who lived in the 1500s.

The goal of much of modern toxicology and pharmacology has since been to define the precise biological factors that dictate the relationship between dose and response following exposure to chemicals. One of the major challenges in risk assessment more recently appreciated is the difficulty in predicting with statistical certainty adverse effects at low doses when the probability of response in a population is low [1]. The nature of the risks at low doses may be qualitatively different from the nature of the risks at high doses. In other words, the risk at low doses is not just a smaller version of the risk at high doses, because the mechanism(s) responsible for the risk at low doses is often different.

Introduction

Paracelsus (1493–1541), or Theophrastus Phillippus Aureolus Bombastus von Hohenheim, is often credited as being the father of modern toxicology for development of the concept of a dose–response relationship. Paracelsus is quoted as saying “all things are poison and nothing is without poison; only the dose makes a thing not a poison” [2]. This principle tells us that even the most innocuous of substances, water for example, can be dangerous in excess. The public is generally aware that dehydration, or an excess of salt relative to water, can be life threatening. However, hyponatremia, an excess of water in the blood, is less widely accepted as a potentially fatal condition. As recently as 2007, a contestant in a radio contest died after consuming 6 l of water in 3 h [3]. The idea that the amount of the exposure or dose determines the frequency or magnitude of response is the fundamental basis for most toxicological research. In theory and practice, the principle that the dose determines the poison is the foundation for experimental toxicology even to this day.

Historically, the relationship between dose and response was considered to be in one direction (i.e., an increase in response with dose or a decrease in response with dose). This single direction, or monotonic, relationship is based on both an understanding of biological processes and mathematical descriptions of experimental outcomes [4]. The most common description of the relationship between dose and response is a sigmoidal curve (Fig. 1). This relationship describes the response expected from the log of the



Ecological and Health Risks at Low Doses. Figure 1

The classical sigmoidal dose–response curve has doses or concentrations that elicit a 50% response (EC50).

Theoretically, there is a threshold concentration or dose below which the response is the same as controls. This spontaneous response is the background rate in the absence of treatment

dose or concentration. Under this relationship, the lowest doses or exposures elicit no measureable response. In general, and in agreement with Paracelsus, organisms have some tolerance for exposure to poisons and toxicants. Some “threshold dose” is required to produce a measureable response. The response then increases exponentially until the midpoint of the curve, where it is half of the maximal response, a point referred to as the EC50 (effective concentration to produce 50% response). The rate of increase in the response decreases above the EC50 to zero where the response reaches its maximum. Increased doses or concentrations above the maximum (EC100) will not increase the response. The sigmoidal curve is the default assumption for the relationship between dose and response.

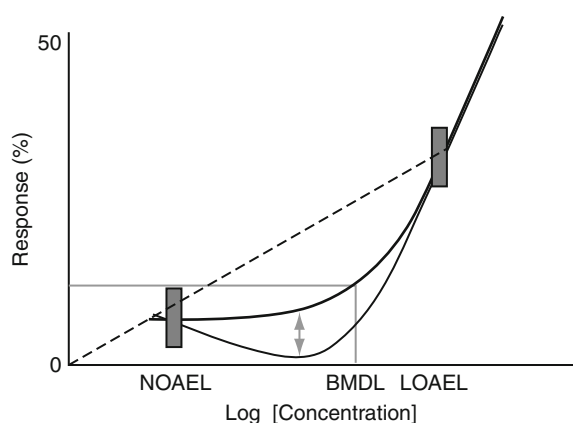
Despite this general convention, agreement on the shape of the dose–response curve at the low end has been controversial [4–6]. The classic sigmoidal dose–response relationship implies that there is a threshold for a biological response. Below that threshold value, exposure to a toxicant should have no measureable effect. Some responses to toxicants, such as the occurrence of tumors, occur spontaneously in control groups. In the absence of exposure, this rate of occurrence is considered the “background” level (Fig. 1). If an incremental dose increases this background rate of

response, then the relationship between the incremental dose and response can be fitted linearly [4]. Furthermore, early studies of the biological basis of tumor formation led to the idea that a single molecule of a carcinogen could cause a mutation that led to tumor development. These “one hit” models have been disputed with both theoretical and physical lines of evidence [4]. For low-energy radiation, theoretical studies indicate that at least two hits of radiation are necessary for mutagenesis, potentially due to high DNA repair rates. Furthermore, cancer research indicates that tumor initiation requires a series of promotional changes even after mutations are incorporated into the genome.

For carcinogens, the precautionary assumption has been that every dose can elicit a response. In other words, there is not an observable no-effect threshold. The most common approach for estimating the low-dose effects of a carcinogen is to adequately describe the relationship between exposure and response over some range of treatments and then extrapolate to lower doses, generally using a linear function. If all poisons have a threshold dose, the linear extrapolation of low-dose response is a conservative assumption. There are reports of DNA adduct formation in proportion to dose even at extremely low concentrations. The US Environmental Protection Agency (EPA) has adopted a policy to utilize the formation of DNA and protein adducts as a means to extend the understanding of the shape of the dose–response relationship beyond the range that can be investigated with cancer bioassays [7]. The relationship between formation of DNA adducts and physiologically meaningful formation of tumors is not as clear. A review of the literature on biomarkers indicates that the dose–response relationship for biomarkers of exposure is not the same as the relationship for biomarkers of effect [7]. Recent descriptions of this debate include the historical progression of acceptance of linear extrapolation by the EPA [8].

In contrast to carcinogens, the assumption of a no-effect threshold has been used extensively for chemicals that are not considered carcinogenic. For these chemicals, the dose at which no adverse effects are detectable is a critical value incorporated into risk assessments. This dose is called the no-observed-adverse-effect level or NOAEL. The NOAEL is derived

from experimental data; it is the highest treatment level which does not differ statistically from the control groups. NOAEL is the basis for regulatory guidance such as acceptable intake levels. Data limitations and experimental design can limit the utility of the NOAEL. For example, note the chosen test concentrations represented by the gray vertical boxes in Fig. 2 for the NOAEL and the lowest observed adverse effects level (LOAEL). The doses in this hypothetical experiment might have been chosen to produce the maximum spread of responses or cover the relevant dose range for this compound. However, a higher concentration may have been chosen which would have still resulted in no significant response – and therefore a different value for the NOAEL. An alternative approach is to use a benchmark dose (BMD) [9]. The BMD is a derived value based on the effective range of the dose–response curve. Calculated BMD have the benefit of using a maximum likelihood approach to determine a confidence interval for the value. The variance in



Ecological and Health Risks at Low Doses. Figure 2 The controversial low end of the dose–response curve is shown here. Experimental data determine the “lowest observed adverse effects level” (LOAEL) and the “no-observed-adverse-effects level” (NOAEL), indicated by the vertical gray boxes. Linear extrapolation from the LOAEL (dashed line) is a conservative estimation of the low-dose effects. Threshold dose can be estimated statistically using a lower bound benchmark dose (BMDL) through a proportion of the response (gray bars). In this hypothetical case, all extrapolations would overestimate toxicity if there were a hormetic response (gray arrow)

the response to the LOAEL or the EC50 can be used to estimate the concentration resulting in 5% response (or another level).

Extensive evidence now indicates that often the relationship between dose and response is biphasic. In other words, low-dose treatments may cause a decrease in response while higher doses cause an increase in response. An extensive review of the toxicological literature [10] found thousands of cases of these types of nonmonotonic dose–response relationships. In many of these cases, low-dose treatments with a toxicant induced a beneficial response – fewer tumors than unexposed controls, for example. A beneficial effect from a low dose of a “poison” or contaminant is referred to as hormesis. These beneficial effects have been reported over a range of responses from growth rate to carcinogenesis. Modeling studies indicate that hormetic response can be driven by a summation of the dose–response relationships for different steps in the biological response, specifically receptor binding, transcriptional activation, rate of DNA repair, and rate of cell division [5]. An example of how multiple physiological responses work together to result in a hormetic response is described below.

Alternatively, low-dose exposure may stimulate an increase in response while higher dose produces a lesser response. This “inverted U” shaped dose–response curve has also been observed in many cases. This relationship between dose and response is characterized by a heightened response to low doses and a depressed response to higher doses. A hypothetical example might be an increase in reproduction observed in response to a low concentration of estrogen and suppressed reproductive rates observed with higher treatments. The more complicated relationships between dose and response demonstrated in cases of hormesis and “inverted U” curves are likely to be observed in systems that require multiple steps to produce the response [5]. Systems biology and genomic/proteomic studies provide an opportunity to test summation hypotheses in a rigorous manner [1, 11]. Modeling and circuitry analysis provide an opportunity to test, or to generate new predictions about the responses of cells, individuals, and populations.

Further complicating the relationship between dose and response is the issue of exposure to mixtures [12]. Some cocontaminants will have the same mechanisms

of action and interact with the same receptors, while others will interact antagonistically through similar mechanisms of uptake, circulation, transformation, or elimination. EPA guidance recommends dose addition for chemicals which have similar mechanisms of action or exhibit response addition. However, predicting the toxicity of mixtures is difficult even when the mechanisms of action are understood.

The focus or target of a research study has a significant impact on predictions related to the dose–response relationship. Many cellular responses are either all or none switches (i.e., cell division). These cellular responses are described as being “quantal,” and the sensitivity of the response in these cases will be described as a cumulative frequency of response at each dose. How the frequency of these incidences relates to specific biological response and adverse health outcomes is largely an unresolved issue [11]. By contrast, examination of a population of cells will produce a smooth dose–response curve as individual cells respond at different concentrations. An analogous dose–response relationship can be described in epidemiological studies [6]. Individuals participating in a study will either have a tumor or not; their response to their particular dose is quantal. However, different individuals will have different sensitivities based on their age, health, lifestyle, and previous exposure history [4]. Each individual is likely to have a different “threshold” for tumor incidence based on their situation. Summation of these different thresholds is expected to “smooth” the dose–response relationship in studies that include a diverse population. Indeed, most human epidemiological studies have a linear relationship between dose and response [6].

Physiological Response to Low Doses

As described above, the response of an organism to a toxicant may be nonmonotonic. One range of treatment may suppress a response while another may enhance the response [12]. Physiological mechanisms underlying these complicated hormetic responses are of much interest. The long history of these types of responses has been documented by Rozman and Doull [13]. Their review summarizes the notion that a beneficial response to a low dose of a poison is the result of homeostatic compensation in response to

a stressor. Rozman and Doull make an analogy between low-dose treatment and physical exercise. Both are minor stressors and homeostatic response to those stimuli may ultimately enhance fitness. Adaptation to a low dose of a stressor, like recovery from physical exercise, may be the result of diverse physiological mechanisms. As a general phenomenon, adaptation or compensation induced by a low-dose treatment is believed to be the basis for hormesis [14].

In situations that result in a hormetic response, continuous exposure to a relatively constant dose should result in a continuous state of (over)compensation. This response has been described as steady-state hormesis induced by chronic exposure [15]. A critical cellular aspect involved in the maintenance of homeostasis is the suite of biotransformation and conjugation enzymes [16]. Three categories of enzymes, Phase I–III, are the basis for xenobiotic metabolism. Phase I enzymes include the cytochrome P450 family, which classically oxidize xenobiotics rendering them more water soluble. However, these Phase I enzymes also activate some procarcinogens and mutagens to reactive metabolites. Phase II enzymes conjugate solubilizing groups that further facilitate excretion. Phase III enzymes are membrane-bound transporters that export metabolized xenobiotics from the cell. A model of these xenobiotic metabolizing enzymes (XMEs) by Zhang and coworkers [17] has generated a mechanistic hypothesis for the hormetic response to many mutagens and carcinogens.

Zhang et al. [17] have described a multipart negative feedback and feedforward model of XMEs, including cross-induction of Phase I and II enzymes. Simulations of this model find a drop in the production of a reactive metabolite below baseline levels in response to 64 h of exposure to an exogenous procarcinogen. Activation of a homeostatic control mechanism, in this case transcriptional induction of XME in response to binding of a nuclear receptor (the aryl hydrocarbon receptor), will compensate for the poison and produce a new physiological state. Because of this hormetic compensation, the authors conclude that the response of biological systems will not be “linearly correlated with the dose of the stressing agent, especially in the low dose region” [17]. The regulation and induction of Phase I and II enzymes is a common cellular mechanism for coping with toxicants. In the cases where Phase I metabolism

activates a toxicant, the feedforward induction of these enzymes is a potential mechanism for hormetic responses at the cellular level. This model provides a hypothesis for the cellular basis of suppression of background levels of tumor induction in response to low-dose exposure to carcinogens to the extent that reactive metabolites are correlated with mutagenesis, apoptosis, and cell proliferation.

Low-level or subacute doses have the potential to produce long-term changes in the magnitude of the response elicited from a toxicant. A decline in response with exposure is called acclimation. The process of acclimation is best studied in relation to exposure of fish to heavy metals (reviewed by Hamilton and Mehrle [18]). Pretreatment of animals with metals, specifically cadmium, copper, mercury, and zinc, is protective for subsequent exposures. Exposure induces synthesis of thionein, a protein that normally occurs only in trace amounts in certain tissues (blood, gills, liver, kidney, and intestine). Binding of metals to this protein, then called metallothionein, reflects the relative concentrations of the metals present and their binding affinity. Chronic toxicity of metals that are sequestered by metallothionein has been described as “spillover”; once metallothionein’s metal binding sites are saturated, the excess will cause tissue damage. This example of competition for a target will influence the shape of the dose–response curve at the low end. Metallothionein is induced by metal, and to a lesser extent, by glucocorticoids. Therefore, some protection or acclimation to metals may be produced by a number of different stressors that produce a glucocorticoid response. In this case, exposure to some classes of metals or heightened stress response will alter the threshold dose for tissue damage from toxicants such as mercury, cadmium, and silver.

Different populations will vary in their response to a toxicant, especially at the low range of the dose–response curve. If a mathematical description of the response indicates that 5% of the population should be impacted, a change in sensitivity of the population (not to mention sample size and health-related factors) will result in more variability in the reported response rate. Populations may differ in age, or individuals may have encountered a previous exposure as described above. Additionally, populations may have a history of chronic exposure that results in transgenerational

changes in the dose–response curve. Genetic or epigenetic adaptation to a chronic stressor will impact the response of a population to exposure. Resistant phenotypes may have enhanced fitness under chronic exposure conditions. Those phenotypes would be expected to increase in frequency in subsequent generations. Experimental studies of killifish from a creosote-contaminated Elizabeth River (Virginia) location have documented resistance to the teratogenic effects of creosote exposure which is maintained through two generations of animals reared in clean conditions [19]. Transgenerational resistance to the polycyclic aromatic hydrocarbons (PAHs), which make up the majority of creosote, is not associated with a general increase in resilience. In fact, offspring of Elizabeth River–collected killifish exhibited higher mortality in response to hypoxic conditions. This study reveals that the magnitude of resistance to PAHs decreases with subsequent laboratory-raised generations (F2 vs. F1); therefore the mechanism of adaptation might not be conferred genetic resistance, but instead be a heritable change in gene regulation.

Epigenetic mechanisms are molecular phenomena, such as DNA methylation of CpG nucleotides, which regulate gene expression without changing the DNA sequence itself [20]. Methylation patterns are normally established two times during development in somatic cells during gastrulation and in the germ cells after sex determination [21]. Many common diseases (e.g., diabetes, osteoporosis, atherosclerosis etc.) have a sex-related incidence bias that may be a result of sex-specific (e.g., X-chromosome inactivation) epigenetic control [22]. As reviewed by Gabory and coworkers, environmental factors including xenobiotics, nutrition, or social factors can influence epigenetic marks throughout development [22]. Furthermore, when germ cells are adversely epigenetically reprogrammed, sex-specific transgenerational effects have been reported. For example, when pregnant rats were exposed to the antiandrogen, vinclozolin, all subsequent generation (F1–F4) males had sperm cell defects and subfertility [21, 23]. Epigenetic changes are also becoming well-recognized mechanisms contributing to cancer and have been associated with all stages of tumor formation and progression [24]. Losses of DNA methylation in oncogenes, hypermethylation of tumor suppressor genes, and altered histone

modifications have all been noted during oncogenesis. The extent to which these epigenetic modifications of gene expression and/or their multigenerational consequences mediate xenobiotic low-dose effects are still largely unknown.

Specific Examples

While a list of xenobiotics that display different effects at high and low doses is likely to grow as “omic” tools and systems biology approaches are implemented in risk assessment, there are a number of examples already known. The bulk of the initial work on predicting low-dose effects was done with respect to carcinogen risk assessment [8]. Radiation and polycyclic aromatic hydrocarbons (PAHs) will be used to highlight issues in cancer low-dose predictions. Similar to cancer progression, neurotoxicological risk assessment has acknowledged that neurological effects are rarely unifactorial [25, 26]. Here, lead and organophosphate pesticides will be used as case studies. A focus of a National Toxicology Program Peer Review [27], certain endocrine disrupting chemicals including bisphenol A (BPA) also have been recognized for their potential to cause “biological change” at low doses. Accordingly, the specific examples described below will focus on carcinogens, neurotoxicants, and endocrine disruptors.

A recent review by Calabrese has provided the history of carcinogen risk assessment [8]. As he describes, the basis for the idea that carcinogen effects were linear at low doses was derived from the early work of Hermann Muller [28], wherein it was discovered that X-ray radiation induced mutations in *Drosophila*. The relationship between low-dose linearity and radiation-induced mutations in germ cells was further asserted in the 1956 Biological Effects of Atomic Radiation (BEAR) Report. The idea that mutations in somatic cells could lead to cancer also with a linear response at low doses was brought to public debate following publication of E.B Lewis’ paper relating radiation risks and leukemia [29]. The criticisms of this manuscript are reviewed by Calabrese [8]; nonetheless, it was a key contribution to the adoption of a precautionary approach and low-dose linear risk assessment for both radiation and ultimately chemical carcinogens. The current risk assessment paradigm for radiation

exposure includes a “single-cell origin, a radiation-induced mutation/aberration as the rate-limiting step, simple expectations of a linear non-threshold radiation response (LNT) at low doses and little or no other relevant contributions. . .to promotion or progression of the tumor” [30]. As pointed out by Goodhead, however, there are a number of radiobiological phenomena including bystander effects and genome instability that are not yet completely understood that may impact how risk of low doses should be assessed and ultimately how human radiation protection is performed. Bystander effects, effects in neighboring nonirradiated cells, are most apparent in the low-dose range where overt toxicity by direct radiation is not dominant [31]. The consequences of bystander effects would be supposed to be relevant to carcinogenicity in that they could be detrimental (toxic effects or mutations in adjacent cells) or beneficial (apoptosis of impacted cells, so no additional tumor progression possible); however, how to mechanistically account for their role in dose–response relationships and risk assessment is currently unknown [30]. Similarly, *de novo* genetic instability in cells generations after irradiation appears to display nonlinear dose responses, and the mechanisms are also unclear. Further research will certainly be necessary to establish the relevance of these more recently recognized potentially low-dose effects of radiation and whether a change in the radiation risk assessment paradigm is necessary [30].

With respect to chemical carcinogens, the Guidelines for Carcinogen Risk Assessment [32] establish a two-step process for dose–response assessment. First, animal or human effect data are used to model biological effects such as adduct, mutation, or tumor formation. The second step then is to extrapolate below the range of observation to the expected range of human exposure with considerations as appropriate or available to mode of action of the chemical, quality of data, and uncertainties therein. Because of the lack of human data, the high costs and low sensitivity of rodent cancer studies, and the goal to protect against one cancer per million, extrapolation of dose back five orders of magnitude is sometimes necessary [33]. Furthermore, it has recently become apparent that biomarkers of exposure (e.g., DNA adducts; linear response) do not have the same dose–response curve shape as biomarkers of effect (e.g., mutations;

threshold response) [7]. As reviewed by Swenberg et al., acrylamide, methymethanesulphonate, and ethylene oxide all represent examples wherein different exposure and effect curves have been reported [7]. Similarly, in a recent 40,800 animal study investigating ultralow-dose effects of the PAH dibenzo[a,l]pyrene in rainbow trout, the initial carcinogen-DNA adduct and the hepatic tumor response curves were “distinctly different” leading the authors to conclude that DNA adducts did not “provide an accurate biomarker for ultralow dose cancer risk evaluation” [33]. Additionally, the study design necessitated significantly less extrapolation below the modeled data set (200-fold vs. typical 100,000-fold) and found that the virtual safe dose for liver tumor formation was 500–1,500 times higher than the standard linear extrapolation from the lower confidence limit on the dose causing 10% incidence typically used in carcinogen risk assessment [33]. Mechanistically, the difference in these curves is relatively easy to appreciate. A specific PAH-DNA adduct will reflect exposure to that parent PAH exclusively. Furthermore, there exist very sensitive analytical methods of DNA adduct measurement. However, it is known that many adducts are not promutagenic. When mutagenic adducts formed, the excretion and repair rates of a particular adduct may be variable because of multiple gene interactions; yet, a relationship between adduct formation and initial exposure will exist. In contrast, mutations are caused by a number of initiating events. There always exists some background of mutations caused by endogenous chemicals or mistakes in DNA replication. Therefore, while physiologically more significant, mutation measurement in exposure experiments will ultimately approach a background frequency at low doses [7]. As these recent reports indicate, the way to measure and predict risks due to exposure to carcinogens is still relevant and continually being reevaluated even though this was one of the first areas in toxicology research to appreciate the need to estimate effects at low doses [8].

Effects of low doses are also important for acute acting compounds. Humans have used and been exposed to lead for thousands of years, and at high doses, a range of adverse effects can occur [34]. In humans, effects depend on developmental stage with the central nervous system as a target in children

and peripheral neuropathies, nephrotoxicity, and cardiovascular effects common in adults [35]. Recent emphasis has been on how childhood low-dose, environmentally relevant lead exposure affects cognitive function. As more data were collected on lead's low-dose toxicity, the regulatory limit of concern for childhood lead intoxication reduced from its initial value of <60 µg/dl set in the 1960s, to 40 µg/dl in 1971, to 30 µg/dl in 1978, and 25 µg/dl in 1985 [36]. The current limit recommended by the Center for Disease Control and World Health Organization is <10 µg/dl. Studies from around the world have found blood lead concentrations far exceeding this level in particular study populations (reviewed in [34, 37]). For example, children from a lead-contaminated area of Ecuador had blood lead concentrations of 52.6 µg/dl (range 9.9–110.0 µg/dl) [38]. Findings from the National Health and Nutrition Examination Surveys (NHANES) in the United States show that average children's blood lead levels have been declining since the 1970s, and the prevalence of elevated levels in children was 0.7% in the 1999–2002 survey [39]. However, recent evidence suggests that blood lead concentrations below 10 µg/dl are not without adverse effect.

The 10 µg/dl standard was based, in part, on a study wherein children were grouped based on umbilical-cord blood concentrations (either <3 µg/dl, between 6 and 7 µg/dl, or >10 µg/dl). At 2 years of age, Mental Development Index scores were significantly lower in children from the >10 µg/dl group [40]. However, using the NHANES III cohort, Lanphear and coworkers found that, after results were corrected for confounding factors, there was an inverse relationship between blood lead concentration and cognitive functioning. Specifically, they reported “for every 1 µg/dl increase in blood lead concentration, there was a 0.7-point decrement in mean arithmetic scores, an approximately 1-point decrement in mean reading scores, a 0.1-point decrement in mean scores on a measure of nonverbal reasoning, and a 0.5-point decrement in mean scores on a measure of short-term memory” [41]. Importantly, these results included children who had blood lead concentrations less than 5 µg/dl. A more recent study by this laboratory followed 172 children measuring their blood lead concentrations at 6, 12, 18, 24, 36, 48, and 60 months of age and tested their performance on the Stanford-Binet

Intelligence Scale at 3 and 5 years of age [42]. They found that as lifetime average blood lead concentrations increased from 1 to 10 $\mu\text{g}/\text{dl}$, IQ declined by 7.4 points using a nonlinear model or 4.6 points with a linear model. Additionally, chelation therapy is largely ineffectual for children with blood lead concentrations between 20 and 44 $\mu\text{g}/\text{dl}$ [43] and cognitive deficits due to lead exposure can last into adulthood [44]. These studies suggest that there is not a clear threshold for lead effects in children.

Unfortunately, the molecular targets and mechanistic reasons to explain low-dose effects of lead exposure are currently lacking. Clearly, critical neurodevelopmental processes occur in the early years of life that have profound effects on learning and memory. A current area of emphasis with respect to mechanisms of lead toxicity is focused on N-methyl-D-aspartate receptor (NMDAR)-mediated synaptic plasticity, specifically hippocampal long-term potentiation [37]. As reviewed by Toscano and Guilarte [37], in rodent models, environmentally relevant lead exposure alters NMDAR subunit expression and function including decreased calcium and nitric oxide signaling, both critical signaling pathways involved in learning and memory. Better understanding of the molecular targets of lead toxicity may in turn offer more effective therapeutic approaches to lead exposure in the future.

Organophosphate pesticides (OP), such as chlorpyrifos, have a well-recognized biomarker of exposure and effect in the irreversible inhibition of acetylcholinesterase (AChE). Significant inhibition of AChE causes a recognizable “cholinergic syndrome” causing accumulation of acetylcholine at cholinergic synapses and overstimulation of cholinergic receptors of the muscarinic and nicotinic type [45]. In addition to the acute cholinergic crisis, high-dose OP exposure can lead to OP-induced delayed polyneuropathy. Percent decreases in plasma pseudocholinesterase or red blood cell cholinesterase (e.g., 30% from baseline) are used in workplace surveillance even though the physiological significance of circulating cholinesterase activity is not known. Furthermore, signs of intoxication typically do not occur until cholinesterase inhibition is $>70\%$ [25]. Accordingly, organophosphate pesticides are another class of neurotoxicant that can exhibit strikingly different low- versus high-dose toxic effects.

Chlorpyrifos is one of the most widely applied OPs. In the United States, registrations for its use in residential settings were canceled in June 2000 [46]. Prior to this ban, chlorpyrifos was detected in 99.7% of personal air samples and 100% of indoor air in an urban pregnant woman cohort [47]. When pregnant rats are given chlorpyrifos at doses below those affecting acetylcholinesterase, their offspring have permanent deficits in learning and memory [25]. Work in the Slotkin laboratory has established persistent changes in serotonergic and dopaminergic systems in adult rats exposed either prenatally or neonatally to chlorpyrifos at doses that are devoid of systemic toxicity [48–50]. Furthermore, in rats exposed neonatally on gestational days 17–20 (during peak neurogenesis), there were sex-dependent behavioral impairments persistent to adulthood [51]. Intriguingly, in addition to the neurotoxicological effects of chlorpyrifos, adverse effects on adenylyl cyclase cellular signaling cascades were noted in heart and liver of developmentally exposed rats [52]. As concluded by Slotkin, “The mechanisms for neurodevelopmental effects from toxins such as chlorpyrifos cannot be presumed to be based on known mechanisms for systemic toxicity in the adult” [25]. More recently, an epidemiologic study of inner-city children compared umbilical cord chlorpyrifos plasma concentrations and cognitive and motor development (at 12, 24, and 36 months of age) and behavior (at 36 months). A significantly higher proportion of children from the “highly exposed” group showed signs of mental and motor delays and were more likely to score in the clinical range for attention problems compared to the lower exposure group at 3 years of age [53]. Further insight into the mechanisms of chlorpyrifos and other organophosphates at low doses may further tighten regulatory control of this class of chemicals and/or offer therapeutic or psychological interventions to remediate low-dose and developmental exposures.

Endocrine disrupting chemicals (EDCs) can change endocrine function and cause adverse effects at the level of the organism, its progeny and/or (sub) populations of organisms. There have been a number of research conferences since the initial Wingspread Conference in 1991, spearheaded by Dr. Theo Colburn, to address the effects of EDCs. Germane to this chapter, in 2000, the

EPA and the National Institute of Environmental Health Sciences (NIEHS) held a peer review “aimed at evaluating the scientific evidence on reported low-dose effects and dose-response relationships for endocrine disrupting chemicals in mammalian species that pertain to assessments of effects on human health” [27]. Nine EDCs, not including dioxins or phthalates esters, were considered. More recently, as reviewed by Vandenberg and coworkers [54], bisphenol A (BPA) individually has been a focus of expert analysis and regulatory review.

Bisphenol A is a high-volume chemical used in polycarbonate plastic. Use of this chemical has recently received intense media scrutiny because of its occurrence in consumer products such as food container liners and baby bottles and ubiquitous occurrence in human samples [55]. Based on relatively low *in vitro* binding affinity to the estrogen receptor compared to estradiol, BPA was initially considered a weak environmental estrogen. However, more recently lower-dose effects are being noted at concentrations below those needed for nuclear estrogen receptor binding [56]. A number of *in vitro* studies measuring BPA-mediated effects in cell types as diverse as adipose explants, and pituitary, prostate cancer, and pancreatic cells have reported U-shaped nonmonotonic dose-response curves (reviewed in [54]). For example, cell proliferation was lower at both low and high doses of BPA compared to intermediate doses. *In vivo*, BPA-mediated effects on CD-1 mouse uterine endpoints also indicated nonmonotonic dose-response curves for some endpoints (e.g., vaginal opening and uterine wet weight) [57]. Fetal exposure to low doses of BPA causes changes in tissue organization and histoarchitecture of developing mammary gland in mice [58, 59]. In males, BPA causes adverse effects on the male reproductive system and sex-related behaviors [60, 61]. Important observations revealed from BPA studies, with particular relevance to low-dose effects, are that experimental design and study controls are critical. For example, compounds present in animal feeds can mask treatment effects, and different animal strains can have different sensitivities [62]. The exact mechanisms underlying BPA low-dose effects are still under investigation. It is important to remember that estrogen is a critical physiological mediator of many diverse tissue-

dependent cellular processes. Furthermore, appropriate timing of sex steroid exposure is key in establishing the hypothalamus-pituitary-gonad feedback necessary for successful development and reproduction. Issues of timing of exposure would be expected to have profound impacts on effects of BPA.

Conclusions from the 2000 peer review included that relevant low-dose effects of BPA included increased prostate weight in male mice and advanced puberty in female mice after *in utero* exposure. Effects in rats (e.g., uterine growth and serum prolactin) were strain specific. They concluded “Data are insufficient to establish the shape of the dose-response curve for BPA in the low dose region, and the mechanism and biological relevance of reported low dose effects are unclear” [27]. In 2006, the NIEHS organized an additional meeting specifically addressing BPA. A product of this meeting was the Chapel Hill Consensus Statement [63], which concluded “. . .that human exposure to BPA is within the range that is predicted to be biologically active in over 95% of people sampled. The wide range of adverse effects of low doses of BPA in laboratory animals exposed both during development and in adulthood is a great cause for concern with regard to the potential for similar adverse effects in humans.” Meanwhile, the National Toxicology Program’s Center for the Evaluation of Risks to Human Reproduction (CERHR) evaluated the scientific data and concluded that some of the previously attributed effects of BPA (e.g., changes in prostate weight, reproductive tract malformations etc.) were not conclusive [64]. This report was highly criticized (see review [54]), and NTP released its own report finally concluding “there is *some concern* for effects on the brain, behavior, and prostate gland in fetuses, infants, and children at current human exposures” but “*minimal concern* for effects on the mammary gland and an earlier age for puberty for females” [65]. Importantly in August 2008, and in an interim update issued January 2010, the US Food and Drug Administration, which regulates food packaging etc., released a draft report finding that BPA remains safe in food contact materials [66]. Some of the questions plaguing the debates on BPA risk assessment include different routes of exposure and metabolism of BPA in animals and humans, lack of consistency in results between laboratory studies, and

the relevancy of the animal models to the human condition. Clearly, this is a timely ongoing debate that absolutely hinges on elucidating the consequences of low-dose environmentally relevant exposure to BPA.

Future Directions

The mechanisms of toxicity and the few selected xenobiotics used as examples here emphasize how timely and relevant the issue of effects at low doses is. Furthermore, answers to the issues raised are key for more informed and appropriate risk assessment. Toxicology studies historically have been limited to testing concentrations representing the high end of environmental relevance. The need to test higher doses was necessitated both by inability to measure/recognize subtle effects, and the need to achieve statistically significant results with the minimal number of test organisms and cost. Now with the realization that, in general, mechanisms of toxicity are not consistent from high to low doses, new robust experiments are needed to explain the more relevant low-dose effects. However, care must be taken so that the cost of regulatory measures that conservatively estimate the harm of low-dose effects is balanced with the cost of potential mechanisms that alter the response of current or future populations. Furthermore, when a biomarker of effect is recognized to not represent true risk of adverse effect (e.g., DNA adduct formation), then more appropriate endpoints (e.g., mutation) should be incorporated into risk assessment. As described above, many low-dose effects are critically linked to the developmental age at the time of exposure. Populations which differ in age or previous exposure are likely to have different responses to low-dose treatment of any contaminant. The World Health Organization has developed a multiple exposure/multiple effects children's health model incorporating the influence of social, economic, and demographic contexts on health outcomes and exposure in addition to the relationship between exposure and effect [67]. As the underlying system biology of test organisms becomes better understood, better evaluation of effects of xenobiotic exposure will be possible [1]. Finally, the field of risk assessment and society in general may ultimately have to deal with the ethical questions that arise associated with potential hormetic effects of chemicals that are toxic at high doses [31].

Bibliography

Primary Literature

1. Andersen ME, Thomas RS, Gaido KW, Conolly RB (2005) Dose-response modeling in reproductive toxicology in the systems biology era. *Reprod Toxicol* 19:327–337
2. Krieger WC (2001) Paracelsus dose response. In: Krieger WC (ed) *Handbook of pesticide toxicology*. Academic, New York
3. Ballantyne C (2007) Strange but true: drinking too much water can kill. *Scientific American* June 21, 2007. <http://www.scientificamerican.com/article.cfm?id=strange-but-true-drinking-too-much-water-can-kill&print=true>. Accessed 24 January 2010
4. Zeise L, Wilson R, Crouch EA (1987) Dose-response relationships for carcinogens: a review. *Environ Health Perspect* 73:259–306
5. Conolly RB, Lutz WK (2004) Nonmonotonic dose-response relationships: mechanistic basis, kinetic modeling, and implications for risk assessment. *Toxicol Sci* 77:151–157
6. Lutz WK (1990) Dose-response relationship and low dose extrapolation in chemical carcinogenesis. *Carcinogenesis* 11:1243–1247
7. Swenberg JA, Fryar-Tita E, Jeong YC, Boysen G, Starr T, Walker VE, Albertini RJ (2008) Biomarkers in toxicology and risk assessment: informing critical dose-response relationships. *Chem Res Toxicol* 21:253–265
8. Calabrese EJ (2009) The road to linearity: why linearity at low doses became the basis for carcinogen risk assessment. *Arch Toxicol* 83:203–225
9. Sand S, Victorin K, Filipsson AF (2008) The current state of knowledge on the use of the benchmark dose concept in risk assessment. *J Appl Toxicol* 28:405–421
10. Calabrese EJ, Baldwin LA (2001) Hormesis: a generalizable and unifying hypothesis. *Crit Rev Toxicol* 31:353–424
11. Andersen ME, Dennison JE, Thomas RS, Conolly RB (2005) New directions in incidence-dose modeling. *Trends Biotechnol* 23:122–127
12. Borgert CJ, Quill TF, McCarty LS, Mason AM (2004) Can mode of action predict mixture toxicity for risk assessment? *Toxicol Appl Pharmacol* 201:85–96
13. Rozman KK, Doull J (2003) Scientific foundations of hormesis. Part 2. Maturation, strengths, limitations, and possible applications in toxicology, pharmacology, and epidemiology. *Crit Rev Toxicol* 33:451–462
14. Stebbing AR (2003) Adaptive responses account for the beta-curve-hormesis is linked to acquired tolerance. *Nonlinearity Biol Toxicol Med* 1:493–511
15. Calabrese EJ, Bachmann KA, Bailer AJ, Bolger PM, Borak J, Cai L, Cedergreen N, Cherian MG, Chiueh CC, Clarkson TW, Cook RR, Diamond DM, Doolittle DJ, Dorato MA, Duke SO, Feinendegen L, Gardner DE, Hart RW, Hastings KL, Hayes AW, Hoffmann GR, Ives JA, Jaworowski Z, Johnson TE, Jonas WB, Kaminski NE, Keller JG, Klaunig JE, Knudsen TB, Kozumbo WJ, Lettieri T, Liu SZ, Maisseu A, Maynard KI, Masoro EJ, McClellan RO,

- Mehendale HM, Mothersill C, Newlin DB, Nigg HN, Oehme FW, Phalen RF, Philbert MA, Rattan SI, Riviere JE, Rodricks J, Sapolsky RM, Scott BR, Seymour C, Sinclair DA, Smith-Sonneborn J, Snow ET, Spear L, Stevenson DE, Thomas Y, Tubiana M, Williams GM, Mattson MP (2007) Biological stress response terminology: integrating the concepts of adaptive response and preconditioning stress within a hormetic dose-response framework. *Toxicol Appl Pharmacol* 222:122–128
16. Parkinson A, Ogilvie BW (2008) Biotransformation of xenobiotics. In: Klaassen CD (ed) Casarett & Doull's toxicology: the basic science of poisons. McGraw Hill, New York, pp 161–304
17. Zhang Q, Pi J, Woods CG, Andersen ME (2009) Phase I to II cross-induction of xenobiotic metabolizing enzymes: a feedforward control mechanism for potential hormetic responses. *Toxicol Appl Pharmacol* 237:345–356
18. Hamilton SJ, Mehrle PM (1986) Metallothionein in fish: review of its importance in assessing stress from metal contaminants. *Trans Am Fish Soc* 115:596–609
19. Meyer JN, DiGiulio RT (2003) Heritable adaptation and fitness costs in killifish (*Fundulus heteroclitus*) inhabiting a polluted estuary. *Ecol Adapt* 13:490–503
20. Jones PA, Takai D (2001) The role of DNA methylation in mammalian epigenetics. *Science* 293:1068–1070
21. Anway MD, Skinner MK (2006) Epigenetic transgenerational actions of endocrine disruptors. *Endocrinology* 147:S43–S49
22. Gabory A, Attig L, Junien C (2009) Sexual dimorphism in environmental epigenetic programming. *Mol Cell Endocrinol* 304:8–18
23. Anway MD, Cupp AS, Uzumcu M, Skinner MK (2005) Epigenetic transgenerational actions of endocrine disruptors and male fertility. *Science* 308:1466–1469
24. Baylin SB, Ohm JE (2006) Epigenetic gene silencing in cancer – a mechanism for early oncogenic pathway addiction? *Nat Rev Cancer* 6:107–116
25. Connors SL, Levitt P, Matthews SG, Slotkin TA, Johnston MV, Kinney HC, Johnson WG, Dailey RM, Zimmerman AW (2008) Fetal mechanisms in neurodevelopmental disorders. *Pediatr Neurol* 38:163–176
26. Weiss B, Cory-Slechta D, Gilbert SG, Mergler D, Miller E, Miller C, Newland MC, Rice D, Schettler T (2008) The new tapestry of risk assessment. *Neurotoxicology* 29:883–890
27. National Toxicology Program (2001) National toxicology program's report of the endocrine disruptors low-dose peer review. NTP Office of Liaison and Scientific Review, Research Triangle Park
28. Muller HJ (1927) Artificial transmutation of the gene. *Science* 66:84–87
29. Lewis EB (1957) Leukemia and ionizing radiation. *Science* 125:965–972
30. Goodhead DT (2010) New radiobiological, radiation risk and radiation protection paradigms. *Mutat Res* 687(1–2):13–16
31. Hoffmann G, Stempsey W (2008) The hormesis concept and risk assessment: are there unique ethical and policy considerations? *Hum Exp Toxicol* 27:613–620
32. EPA US (2005) Guidelines for carcinogen risk assessment. US EPA, Washington DC
33. Bailey GS, Reddy AP, Pereira CB, Harttig U, Baird W, Spitsbergen JM, Hendricks JD, Orner GA, Williams DE, Swenberg JA (2009) Nonlinear cancer response at ultralow dose: a 40800-animal ED (001) tumor and biomarker study. *Chem Res Toxicol* 22:1264–1276
34. Tong S, von Schirnding YE, Prapamontol T (2000) Environmental lead exposure: a public health problem of global dimensions. *Bull World Health Organ* 78:1068–1077
35. Liu J, Goyer RA, Waalkes MP (2008) Toxic effects of metals. In: Klaassen CD (ed) Casarett and Doull's toxicology: the basic science of poisons. McGraw Hill, New York, pp 931–979
36. Landrigan PJ (2000) Pediatric lead poisoning: is there a threshold? *Public Health Rep* 115:530–531
37. Toscano CD, Guilarte TR (2005) Lead neurotoxicity: from exposure to molecular effects. *Brain Res Brain Res Rev* 49:529–554
38. Counter SA, Vahter M, Laurell G, Buchanan LH, Ortega F, Skerfving S (1997) High lead exposure and auditory sensory-neural function in Andean children. *Environ Health Perspect* 105:522–526
39. CDC (2005) Blood lead levels – United States, 1999–2002. *MMWR* 54:513–516
40. Bellinger D, Leviton A, Waternaux C, Needleman H, Rabinowitz M (1987) Longitudinal analyses of prenatal and postnatal lead exposure and early cognitive development. *N Engl J Med* 316:1037–1043
41. Lanphear BP, Dietrich K, Auinger P, Cox C (2000) Cognitive deficits associated with blood lead concentrations <10 microg/dL in US children and adolescents. *Public Health Rep* 115:521–529
42. Canfield RL, Henderson CR Jr, Cory-Slechta DA, Cox C, Jusko TA, Lanphear BP (2003) Intellectual impairment in children with blood lead concentrations below 10 microg per deciliter. *N Engl J Med* 348:1517–1526
43. Rogan WJ, Dietrich KN, Ware JH, Dockery DW, Salganik M, Radcliffe J, Jones RL, Ragan NB, Chisolm JJ Jr, Rhoads GG (2001) The effect of chelation therapy with succimer on neuropsychological development in children exposed to lead. *N Engl J Med* 344:1421–1426
44. White RF, Diamond R, Proctor S, Morey C, Hu H (1993) Residual cognitive deficits 50 years after lead poisoning during childhood. *Br J Ind Med* 50:613–622
45. Costa LG (2008) Toxic effects of pesticides. In: Klaassen CD (ed) Casarett & Doull's toxicology: the basic science of poisons. McGraw Hill, New York, pp 883–930
46. EPA US (2002) Interim reregistration eligibility decision for chlorpyrifos. US EPA, Washington DC, pp 1–235
47. Whyatt RM, Barr DB, Camann DE, Kinney PL, Barr JR, Andrews HF, Hoepner LA, Garfinkel R, Hazi Y, Reyes A, Ramirez J, Cosme Y, Perera FP (2003) Contemporary-use pesticides in personal air samples during pregnancy and blood samples at delivery among urban minority mothers and newborns. *Environ Health Perspect* 111:749–756

48. Aldridge JE, Meyer A, Seidler FJ, Slotkin TA (2005) Alterations in central nervous system serotonergic and dopaminergic synaptic activity in adulthood after prenatal or neonatal chlorpyrifos exposure. *Environ Health Perspect* 113:1027–1031
49. Roy TS, Sharma V, Seidler FJ, Slotkin TA (2005) Quantitative morphological assessment reveals neuronal and glial deficits in hippocampus after a brief subtoxic exposure to chlorpyrifos in neonatal rats. *Brain Res Dev Brain Res* 155:71–80
50. Slotkin TA, Seidler FJ (2005) The alterations in CNS serotonergic mechanisms caused by neonatal chlorpyrifos exposure are permanent. *Brain Res Dev Brain Res* 158:115–119
51. Levin ED, Addy N, Baruah A, Elias A, Christopher NC, Seidler FJ, Slotkin TA (2002) Prenatal chlorpyrifos exposure in rats causes persistent behavioral alterations. *Neurotoxicol Teratol* 24:733–741
52. Meyer A, Seidler FJ, Slotkin TA (2004) Developmental effects of chlorpyrifos extend beyond neurotoxicity: critical periods for immediate and delayed-onset effects on cardiac and hepatic cell signaling. *Environ Health Perspect* 112:170–178
53. Rauh VA, Garfinkel R, Perera FP, Andrews HF, Hoepner L, Barr DB, Whitehead R, Tang D, Whyatt RW (2006) Impact of prenatal chlorpyrifos exposure on neurodevelopment in the first 3 years of life among inner-city children. *Pediatrics* 118: e1845–e1859
54. Vandenberg LN, Maffini MV, Sonnenschein C, Rubin BS, Soto AM (2009) Bisphenol-A and the great divide: a review of controversies in the field of endocrine disruption. *Endocr Rev* 30:75–95
55. Calafat AM, Ye X, Wong LY, Reidy JA, Needham LL (2008) Exposure of the US population to bisphenol A and 4-tertiary-octylphenol: 2003–2004. *Environ Health Perspect* 116:39–44
56. Welshons WV, Nagel SC, vom Saal FS (2006) Large effects from small exposures III. Endocrine mechanisms mediating effects of bisphenol A at levels of human exposure. *Endocrinology* 147:S56–S69
57. Markey CM, Michaelson CL, Veson EC, Sonnenschein C, Soto AM (2001) The mouse uterotrophic assay: a reevaluation of its validity in assessing the estrogenicity of bisphenol A. *Environ Health Perspect* 109:55–60
58. Markey CM, Luque EH, De Munoz TM, Sonnenschein C, Soto AM (2001) In utero exposure to bisphenol A alters the development and tissue organization of the mouse mammary gland. *Biol Reprod* 65:1215–1223
59. Vandenberg LN, Maffini MV, Wadia PR, Sonnenschein C, Rubin BS, Soto AM (2007) Exposure to environmentally relevant doses of the xenoestrogen bisphenol-A alters development of the fetal mouse mammary gland. *Endocrinology* 148:116–127
60. Farabollini F, Porrini S, Della SD, Bianchi F, Dessi-Fulgheri F (2002) Effects of perinatal exposure to bisphenol A on socio-sexual behavior of female and male rats. *Environ Health Perspect* 110(Suppl 3):409–414
61. Richter CA, Birnbaum LS, Farabollini F, Newbold RR, Rubin BS, Talsness CE, Vandenberg JG, Walser-Kuntz DR, vom Saal FS (2007) In vivo effects of bisphenol A in laboratory rodent studies. *Reprod Toxicol* 24:199–224
62. Vom Saal FS, Richter CA, Ruhlen RR, Nagel SC, Timms BG, Welshons WV (2005) The importance of appropriate controls, animal feed, and animal models in interpreting results from low-dose studies of bisphenol A. *Birth Defects Res A Clin Mol Teratol* 73:140–145
63. vom Saal FS, Akingbemi BT, Belcher SM, Birnbaum LS, Crain DA, Eriksen M, Farabollini F, Guillette LJ Jr, Hauser R, Heindel JJ, Ho SM, Hunt PA, Iguchi T, Jobling S, Kanno J, Keri RA, Knudsen KE, Laufer H, LeBlanc GA, Marcus M, McLachlan JA, Myers JP, Nadal A, Newbold RR, Olea N, Prins GS, Richter CA, Rubin BS, Sonnenschein C, Soto AM, Talsness CE, Vandenberg JG, Vandenberg LN, Walser-Kuntz DR, Watson CS, Welshons WV, Wetherill Y, Zoeller RT (2007) Chapel Hill bisphenol A expert panel consensus statement: integration of mechanisms, effects in animals and potential to impact human health at current levels of exposure. *Reprod Toxicol* 24:131–138
64. National Toxicology Program (2007) NTP-CERHR expert panel report on the reproductive and developmental toxicity of bisphenol A. In: Center for the Evaluation of Risks to Human Reproduction (ed) National toxicology program, pp 1–396
65. National Toxicology Program (2008) NTP-CERHR monograph on the potential human reproductive and developmental effects of bisphenol A. In: Center for the Evaluation of Risks to Human Reproduction (ed) National toxicology program, pp 1–321
66. US FDA (2010) Update on Bisphenol A for use in food contact applications. In: FDA (ed) US food and drug administration, pp 1–7
67. Briggs D (2003) Making a difference: indicators to improve children's environmental health. World Health Organization, Geneva, pp 1–17

Ecological Risk Assessment and Animal Models

LISA J. BAIN

Department of Biological Sciences, Clemson University, Clemson, SC, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Animals Models in the Basic Framework of Risk Assessment

Considerations When Determining Which Organisms to Use for an Ecological Risk Assessment

Field Studies and Toxicity Testing
 Other Assays Providing Support to Risk Assessments
 Future Directions Using Animal Models in Risk Assessments
 Bibliography

Glossary

Bioaccumulation factor (BAF) The concentration of contaminant in the organism divided by the concentration of the contaminant from all exposure sources, including water, air, soil, sediment, or food.

Bioconcentration factor (BCF) The concentration of contaminant in the organism divided by the concentration of the contaminant in the water.

Biomarkers A quantifiable change in a biochemical, physiological, or histological parameter that can be used as a measure of response to a toxicant.

Biological survey A systemic gathering of information on the presence, abundance, and condition of species at a given site to assess the health of those organisms.

EPT index The number of organisms in the orders *Ephemeroptera* (mayflies), *Plecoptera* (stoneflies), and *Trichoptera* (caddisflies). These orders are typically considered to be sensitive to contaminants.

LC₅₀ The median lethal concentration, or the concentration of the medium that was lethal to 50% of the test organisms.

LOEC The lowest observed effect concentration, or the lowest concentration of test medium in which statistically significant toxic effects are seen.

NOEC No observed effect concentration, or the highest concentration of test medium in which no significant toxic effects are seen.

Toxicity test An assay in which organisms are exposed to soil, water, or air, and the effects of contaminants on processes such as survival, growth, and reproduction are determined.

Toxicogenomics Examining changes in gene expression, protein expression, and/or metabolite profiles in a response to an exposure by a toxicant.

Definition of the Subject

An ecological risk assessment attempts to determine the effects of contaminants or other environmental stressors on individuals, populations, and communities of species. Although ecological risk assessments can

be used for regulating chemicals and managing watersheds, the most typical use is for the study and remediation of contaminated sites. The organismal effects that can be examined vary considerably, depending on the ultimate use of the assessment, but most often focus on the organisms' ability to survive, grow and develop appropriately, and reproduce. The choice of which organisms to use should be based on easily evaluated criteria, such as the ease of sampling a particular organism, ensuring that the organism can both respond to the toxicant or mixture of toxicants, and that the organismal response is graded with respect to increasing exposure concentrations and/or times. The specific endpoints measured in the risk assessment with respect to organisms can therefore include lethality, weight gain, population numbers, birth and death rates, alterations in community structure, among others. Some endpoints are easily quantified, such as lethality. Other endpoints may be much more difficult to measure, such as reproduction, development, productivity, and immune function. For these more subtle outcomes, one often uses biological markers to indirectly assess these endpoints, or surrogate species in the laboratory to find out more about the health of the population or community under study. Other methods to better refine the risk assessment include the use of models and the use of genomics. The underlying goal to conduct an effective assessment relies on the choice of the organism and the endpoint examined so that management decision is appropriate and protective of organismal health.

Introduction

The selection of which organisms to monitor or use is an important part of the risk assessment process. This entry first discusses the role of organismal selection and response in the basic framework of risk assessment. The criteria used when selecting an appropriate organism to use and the common organisms routinely used in assessments will be discussed. Laboratory toxicity tests, using either standard organisms as surrogate species or nontraditional organisms that are more representative of the site can be used. Field studies are also routinely conducted, which can monitor the presence and the abundance of a variety of species. Beyond typical field assessment, biotic indices are used as

measures of community structure, while biomarker data are used to assess sublethal impacts that often cannot be determined during standard laboratory toxicity testing. Finally, the entry will cover future directions in risk assessment, which may include the use of genomics, proteomics, other multi-end point analysis, and physiologically-based pharmacokinetic (PBPK) modeling.

Animal Models in the Basic Framework of Risk Assessment

Overview

The basic framework for risk assessment consists of four phases: problem formulation, analysis, risk characterization, and risk management [1, 2]. Much of this entry will focus on the first two phases, where the selection and analysis of organismal responses is most crucial. There is an additional step prior to the risk assessment itself, which is termed planning [3]. At the planning stage, the risk manager, risk assessors, and stakeholders should work together to identify the goals of the risk assessment. These goals should include developing clear objectives for what organisms the manager is most interested in protecting, determining what contaminants are most important to consider for the remediation effort, and indicating the acceptable levels of contaminants or the desired condition of the environment and organismal health after the remediation is completed [4].

During the problem formulation stage, a conceptual model should be developed to describe the relationships between the contaminants at the site and the organisms that may be affected by them. Although many species can be found at a site, the risk assessment is only concerned with those that could be adversely affected. Additionally, since there are always factors that mitigate or exacerbate the concentration of chemical that an organism encounters, these should be taken into account during model development. The risk assessor should identify the key ecological characteristics that allow or prohibit certain species to thrive. Potential confounding factors should be identified, such as physical manipulations, building and construction, other human disturbances, or competition by

invasive species [4]. There is also the issue of indirect effects, in which the impact on a species affects another. Examples include a loss of a prey or predator species, or loss of vegetation that provides food and shelter. Once these factors are taken into account, then appropriate selection of the animal models to use in the risk assessment can occur.

During the analysis stage, there is both a characterization of the exposure and a characterization of the effects. Organisms may be sampled from the exposed site and changes due to the exposure determined. If suitable native species cannot be found or cannot be found in great enough numbers, outside organisms may be brought in and caged at the site to determine the amount of toxicant being taken into the organism and its resultant effects. More frequently, if a suitable field study cannot be performed on site, laboratory tests are conducted using surrogate species in standardized tests to determine the effects of contaminant exposure.

After data collection, one must determine whether there is a relationship between the contaminants and effects on organisms. One useful way to categorize the strength of the evidence is based upon the Bradford-Hill criteria [5], which provide a basic framework for assessing the strength of causal relationships. Although Hill originally published these criteria for human epidemiological work to determine the causes of diseases, they have been modified for ecotoxicology and risk assessments [6, 7]. The criteria or types of evidence a risk assessor is looking for to support a cause-effect relationship include:

1. *Spatial co-occurrence* – the effect should be observed only where the contaminant is present.
2. *Temporality* – there should be a cause-effect time relationship, in that the effect has to occur after the exposure to the contaminant.
3. *Biological Gradient* – this criterion describes the classical exposure-response relationship, in which higher exposure concentrations should lead to a greater incidence or more severe effects if the effect is due to the contaminant.
4. *Complete Exposure Pathway* – the contaminant has to be able to reach the organism in order to cause an effect. If the toxicant is in media that are

inaccessible to the organism of interest, than exposure will not occur.

5. *Consistency of Association* – if similar observations are documented by multiple investigators in multiple locations, this increases the likelihood that the effects seen are due to the contaminant.
6. *Experiment* – if the exposure is manipulated or modified, this should cause a decrease in effects. The manipulation can be conducted by using site media in the laboratory and altering the contaminant's concentration by dilution, or by conducting field experiments in which the exposure can be modified, contained, or eliminated.
7. *Plausibility* – this criterion asks whether the effects seen in the organisms can be explained because of a known or plausible biological mechanism. Additionally, the risk assessor should understand whether the effects would be expected at the concentration of contaminant seen at the site.
8. *Analogy* – this criterion examines uses case studies from contaminants that are similar in structure to the suspected contaminant, and asks whether the expected effects be the same.
9. *Specificity of Cause* – is the observed effect seen consistently associated with the contaminant or likely to only be caused by the contaminant. If so, the cause-effect relationship is quite strong.
10. *Predictive Performance* – this criterion asks whether side effects occurred as a result of the exposure, and if so, could they have been predicted based upon the contaminant's mode of action.

The first four provide the strongest evidence, since they are directly based upon the site being assessed. The next six are often considered less strong, since the data for these criteria typically come from laboratory studies or other field-based studies, which may or may not be completely applicable for the site in question. However, they provide corroborating evidence that, when coupled with the first four, can support the risk characterization process.

Monitoring Trends in the Environment

Knowing the baseline levels of contaminants and their impact on the health of populations and communities can support or act as a precursor to a risk assessment.

Baseline levels are found by monitoring trends in the environment. Although these are not risk assessments, they can provide valuable information to the risk assessor. The US EPA's Environmental Monitoring and Assessment Program (EMAP) is one such large-scale endeavor that seeks to monitor sites in a particular region for many years, thus providing data on multiple spatial and temporal scales [8, 9]. EMAP data can include such information as contaminant levels, bacterial loads, and species distributions. Thus, the data can confirm trends in a particular area or provide baseline levels to serve as a reference site to aid risk assessors.

Biological surveys may also provide valuable information to support a risk assessment. Biological surveys most often are conducted in aquatic systems and often have data bases that span multiple sites and years. They are typically used to determine if an aquatic system is meeting its designated aquatic life uses, but have also been used to evaluate the effectiveness of controls and restoration activities, and characterize regional biotic attributes of reference sites. These surveys determine the presence or abundance of species at a given point in time. Information may also be collected on the numbers of sensitive or tolerant species, and may provide information on the overall community composition. These data is often converted and used in as a biotic index, which assigns numbers to analyzed sites to indicate the level of impact that external factors have on the community [10]. The survey can provide baseline data on the health of a particular system or region, which can be used as a reference during the risk assessment to determine whether the impacted site is similar or not.

Ecological epidemiology is another method to monitor trends in the environment. This is often used prior to a risk assessment, when the sources of contamination are either unknown or due to nonpoint sources [11]. Most often, this endeavor is initiated after reports of a noticeable environmental effect, such as the mortality of a large number of individuals, an increase in the number of deformities, or the decline of a species. The goal of the epidemiological study is to try to form a causal chain from source to exposure to effects. Once the cause is identified, those parties involved can then formulate the problem and continue on with the traditional risk assessment.

Considerations When Determining Which Organisms to Use for an Ecological Risk Assessment

The basic purpose of the risk assessment is either to estimate the exposure to the consumers or to the end-point organisms, such as a carnivore higher up the food chain. Therefore, the relevance of the assessment heavily depends on choosing the appropriate organisms to monitor or study, or the appropriate species assemblages. There are certain criteria that are desirable when considering which organisms to use in the risk assessment [4]. First, the organism should be easily sampled. This suggests that the organism should be abundant in the area under study and amenable to capture and holding. Second, the organism should be sensitive to the contaminants in the system. If a particular species is known to be insensitive to PAHs, for example, than selecting it to determine effects of creosote contamination will not be protective of the environment. Third, the organism should respond to the toxicant under question in a predictable manner. Ideally, one wants a gradual change in response when the contaminant levels are increased. Ideally, a threshold should exist in which contaminant concentrations below these levels should be of little concern because they will not cause adverse effects [4, 12].

Levels of Biological Organization

There are different levels of biological organization that can be used in an ecological risk assessment: (1) monitoring or testing individual organisms; (2) testing or monitoring populations of a particular species; or (3) community level assessment (Table 1). The choice of level will ultimately depend upon a variety of factors, such as the types of organisms present, their abundance, and their types of interactions with one another. Using the assessment to investigate impacts on a rare species might necessitate only examining a few individual organisms. However, if during the problem formulation step, there is concern about indirect effects, in which the impact on one species affects another, than population or community levels may be most appropriate. For example, if the contaminant is thought to reduce levels of a prey species, then using a community level approach to determine predator–prey interactions might be in order.

Ecological Risk Assessment and Animal Models.

Table 1 Levels of biological organization and typical end-points assessed during a risk assessment

Organism	Population	Community
Survival	Survival	Survival
Growth	Population density	Population density
Condition factor	Reproductive success	Community structure changes
Reproductive success	Productivity	Trophic imbalances
Tumors	Changes in genetic structure	Biomass
Behavioral alterations	Development of resistance	
Acetylcholinesterase inhibition	Biomass	
Stress proteins		
Changes in metabolism		
Immune system alterations		

The spatial characteristics of the site and the specific agent or agents being assessed will clearly influence the choice of biological organization, as will any temporal variations in the exposure. For example, if one is doing an assessment involving an effluent stream released once per month, this will necessitate a different strategy for testing and monitoring than an assessment for remediation of a contaminated site in which the organisms are continuously exposed. Along with temporal variations, the concentration of the agents may make a difference in the choice of scale as well. An exposure that is extremely high in concentration and pulsatile in nature would warrant different tests than a low level, chronic exposure [3].

The first level of organization is to monitor or test individual organisms. The endpoints garnered from these tests are typically survival, growth, and/or reproduction (Table 1). Depending on the organism used, these tests can take anywhere from hours to days. Population level tests typically are conducted via microcosms or mesocosms, using multiple organisms.

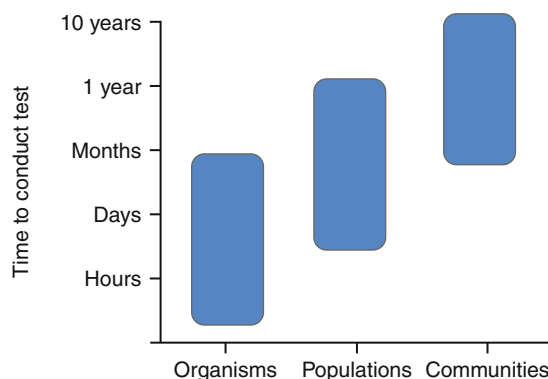
The endpoints assessed are usually survival and growth/production, which takes days to months to conduct. The effects on populations can also be determined in the field by either a designed test or, more often, by simply monitoring the environment. Examples of these types of investigations might include examining organisms in a lake or pond, or in gridded and staked terrestrial areas. Again, the endpoints often assessed are survival, growth, reproduction, and production, which may take days to years to conduct. Community assessments are conducted either using a mesocosm (a small community), field tests, or environmental monitoring. Beyond the typical endpoints of survival and growth, one can now examine changes in community structure and trophic balance.

Exposure Time, Organismal Characteristics, and Exposure Routes

Once the level of biological organization is decided upon, the next step is to determine the types organism(s) to use and how long to expose or monitor them. Influencing factors for the exposure time and type of organism include the type of media that are contaminated (soil, sediment, water), the concentrations of the contaminants, and what types of organisms are present at the site, and how many of them are at the site.

The length of exposure time for each of these tests depends on the agent and how it is released or occurs in the area. If assessing the effects of a pesticide application, one may only need to conduct the test or monitor the environment for a few days. If assessing the effects of an oil spill, monitoring for months to years might be more appropriate (Fig. 1). The timing of when the test is conducted also needs to take into account issues like seasonal variability. Water and food may be limited at certain times during the year, and thus the organisms' exposure could be reduced. Additionally, breeding and migration times may pose issues for accurate monitoring and should be taken into account before beginning the study.

Most risk assessments focus on a limited number of organisms – typically no more than just a few species or a small community, if looking at invertebrates or algae. The choice of organisms to examine therefore becomes critical. Among the specific factors and organismal characteristics that should be taken into consideration



Ecological Risk Assessment and Animal Models. Figure 1 Time scale for collecting field or laboratory data to support the risk assessment

is the mobility of the organism. Ideally, one would want to monitor an organism with limited mobility, such as earthworms or invertebrates, or to monitor those organisms with low mobility at some stage in their life cycle [12]. Another concern is the sensitivity of species toward the contaminant of interest. Ideally, one would want to use the most sensitive species, which would be protective of all other species. However, given that most sites have a suite of contaminants, finding the species most sensitive to the mixture may be difficult. A third consideration is the ecological function or role of the organism. Often, it is good to choose at least one primary producers to study because of their large contribution to the food chain. Other factors to take into account are choosing an organism that is easy to collect and choosing an organism of high relative abundance. Finally, one also needs to be concerned with a species that has a certain recognized value, either economic or symbolic, as the one that you want to protect. It may be necessary to monitor these types of species, even if they do not fit the other criteria for choice of an organism to use [12].

The last issue is to determine the most likely route of uptake, be it oral, dermal, or inhalation (for a recent review, see [13]). The oral route is the most likely exposure route (Table 2). However, organisms can also swim in water, absorb contaminants through dermal contact, and take up compounds via inhalation, so that depending upon the contaminants at the site, these minor routes of uptake may be more important. For terrestrial invertebrates, the most common type of

Ecological Risk Assessment and Animal Models.

Table 2 Routes of uptake in different organisms

Organism	Common uptake routes
Terrestrial invertebrates	Oral, dermal
Aquatic invertebrates	Oral, respiratory surfaces
Fish	Oral, gills, dermal
Mammals	Oral >> dermal, inhalation
Birds	Oral, inhalation
Amphibians	Dermal, oral
Reptiles	Oral

Routes listed are in the order of importance for each class of organism

organism to monitor is the earthworm. For these animals, exposure occurs either dermally or orally. Other arthropods can be examined, but these typically have an exoskeleton, so a food-borne exposure is more likely than dermal exposure. Aquatic invertebrates are most likely exposed through the oral route. Rather than conducting a test or monitoring these organisms, it is sometimes more desirable to use water or sediment bioaccumulation factors (BAFs) to estimate exposure. The uptake route of fish is typically oral but can also be via the gills or as a dermal exposure. As with aquatic invertebrates, one sometimes estimates the exposure to fish using BAFs or bioconcentration factors (BCFs). BAFs tend to be more consistent across different water chemistries and different organisms. They are also considered more realistic of an exposure since diet and food chain multipliers are included within the BAF. BCFs are determined in laboratory-based studies and are considered more precise. Therefore, a BCF is often what is used to estimate exposure, if available [3]. Exposure factors for many organisms can be found in the *Wildlife Exposure Factors Handbook* [14].

For most terrestrial organisms, the oral uptake route tends to be most important (for a recent review, see [13]). In mammals, the uptake of contaminants tends to be through the diet, while the dermal and inhalation routes tend to be negligible. Mammals have extremely varied feeding strategies, both in terms of prey preference and in feeding frequency. This can result in high exposures to some mammals while no exposure to others, so care must be taken in the risk

assessment to choose a species that will come into contact with the contaminants of concern. Periods of fasting, hibernation, and migration can mobilize stored contaminants, resulting in toxicity without changing the body burden. Additionally, lactational and placental transfer may occur from mothers to the developing offspring. In birds, the oral route is the most important uptake pathway, most commonly through the ingestion of contaminated prey items. During pesticide exposure, uptake via inhalation may contribute significantly to their body burdens. The transfer of contaminants from gravid females into their eggs may also represent a significant exposure source to embryos. In reptiles, the oral route is most important via ingestion of contaminated food items. The skin tends to inhibit some compounds from passing through dermally. As with avian species, contaminant transfer into eggs also represents an important exposure route. In contrast, the dermal route may be extremely important to amphibians due to their life cycle, which is spent in contact with both water and soil for parts of their life span.

When quantifying how much contaminant the organism had received during the exposure, one can determine its concentration in the whole body of the organism, its burden in an organ in which the toxicant is likely to exert a toxic effect, such as the liver, kidney, or brain, or in organs of accumulation, such as fat, hair, or feathers. As stated above, for aquatic organisms, BAFs or BCFs are often used to estimate exposure.

Field Studies and Toxicity Testing

Toxicity tests involve collecting contaminated soil, sediment, or water from the site, and then exposing organisms to that medium or those media in a laboratory setting. There are several different advantages of toxicity tests over field studies. First, a toxicity test can evaluate whether the contaminant was bioavailable to the organisms. Second, a toxicity test can enable a more informed assessment of the impact of mixtures of compounds than the use of data in the literature on single compounds. They can determine whether characteristics of the medium itself, such as hardness or pH, are influencing toxicity. Third, they can examine substances whose effects on wildlife are not well understood. Fourth, they are typically used for monitoring

to determine whether the risk mitigation strategy is working [2, 15].

Although a variety of endpoints can be measured from toxicity tests, the most common ones are survival, growth, and reproduction. Survival is most often assessed using a short term, acute toxicity test. An acute test typically lasts 24–96 h, during which the organisms are exposed to a medium from the site, plus a reference medium, and the number of deaths is recorded. Outcomes of this test include the medium lethal concentration or LC_{50} , which is the concentration of the medium that killed 50% of the test organisms. Other values that are typically obtained are the lowest observed effect concentration (LOEC), which is the lowest concentration of test medium in which statistically significant toxic effects are seen, and the no observed effect concentration (NOEC), which is the highest concentration of test medium in which no significant toxic effects are seen.

Sublethal endpoints, such as growth and reproduction, are best measured using a chronic toxicity test. These tests take a longer period of time to conduct, because they encompass either a critical developmental phase or the organism's entire life cycle, but give better resolution as to whether a toxic effect is occurring on site. Similar to an acute test, the values obtained are typically a median effective concentration (EC_{50}), or the concentration of medium that caused effects, such as reduced growth or reproduction, on 50% of the organisms. A LOEC and NOEC are typically calculated for chronic tests.

Generally, values such as the LC_{50} , NOEC, and LOEC for a single contaminant using standardized test organisms can be easily be obtained from the literature or using databases such as the EPA's ECOTOX database (<http://cfpub.epa.gov/ecotox/>) or the USGS's Contaminant Exposure and Effects-Terrestrial Vertebrates database (CEE-TV) (<http://www.pwrc.usgs.gov/contaminants-online/pages/ceetv/ceetvintro.htm>). This means that oftentimes, a *de novo* test may not be necessary. However, one should consider conducting a toxicity test if the data for the compound or interest is absent in the literature, if there is reason to suspect that conditions at the site, such as pH or organic matter, may be influencing bioavailability or toxicity to organisms, or if the interactions between mixtures of compounds are considered significant.

Standard Toxicity Tests

When the risk assessor decides that a *de novo* toxicity test is needed, there are several elements that need to be decided upon. First, the assessor should decide upon the level of effort needed to achieve the goals of the study. If he or she only wants to know whether the medium is toxic or not, using two test organisms and collecting medium from just a few sites may be sufficient to obtain the desired information. However, to fully characterize the site's toxicity, the study might use test organisms at different trophic levels, a number of sampling locations, and several dilutions of each test medium [15]. Second, a reference site should be included that is similar in characteristics to the site under study and is located as close to the study site as possible. Third, the test medium must be decided upon. Although it is most typical to use the water, sediment, or soil directly from the site, one can use soil or sediment elutriate to expose the test organisms. Finally, the test organisms themselves need to be selected. Whenever possible, a standardized organism should be used that is representative of the onsite organisms and is sensitive to the contaminant of concern [15, 16].

Several types of acute aquatic tests involving fish, macroinvertebrates, mysids, *Daphnia*, and algae have been developed (for a good overview, see [17]). Most of these are 48–96 h LC_{50} tests that measure survival as the endpoint (Table 3). Chronic tests generally involve either early life stage tests with juveniles, but there are some partial or full life cycle tests. Most of these measure endpoints such as survival, growth, and reproduction (Table 3).

Nonstandard Tests and Field Tests

If a laboratory test is needed to assess site medium, the risk assessor should try to choose a standardized test and a standardized organism (Table 3). This is not always possible, as the biota in a particular site may not be similar enough to a standardized organism, or the standardized organisms are not sensitive to the contaminant at the site. In these cases, using an alternative species is completely appropriate [2, 15]. Additionally, field studies are routinely conducted at the actual site. These studies are more concerned with effects on populations and communities, rather than

Ecological Risk Assessment and Animal Models. Table 3 Standard toxicity tests

Test	Organism	Test duration	Endpoints	Reference
Freshwater acute	<i>Daphnia magna</i> or <i>Daphnia pulex</i>	48–96 h	Survival	[18, 19]
	Fathead minnow	96 h	Survival	[18, 19]
	Rainbow trout	96 h	Survival	[18, 19]
	Brook trout	96 h	Survival	[18, 19]
Freshwater chronic	<i>Cerodaphnia dubia</i>	7 days	Survival, reproduction	[20]
	<i>Daphnia magna</i>	21 days	Survival, reproduction	[21]
	Fathead minnow larvae	7 days	Survival, growth	[22]
	Algae (<i>Selenastrum capricornutum</i>)	96 h	Cell density	[23]
Saltwater acute	Bivalve mollusk larvae	48 h	Abnormal shell development	[24]
	Silverside (<i>Menidia</i> sp)	48 h	Survival	[18, 19]
	Sheepshead minnow (<i>Cyprinodon variegates</i>)	48 h	Survival	[18, 19]
Saltwater chronic	Silverside larvae	7 days	Survival, growth	[18, 19]
	Sheepshead minnow	7 days	Survival, growth	[18, 19]
	Mysid shrimp	7 days	Survival, growth, reproduction	[25]
	Sea urchin (<i>Arbacia punctulata</i>)	20 min	Egg fertilization	[25]
	Algae (<i>Champia parvula</i>)	5–7 days	Fertilization	[26]
Freshwater sediment chronic	<i>Hyalella azteca</i>	10–30 days	Survival, growth, reproduction	[26]
	Midge (<i>Chironomid tetans</i> or <i>C. riparius</i>)	10–14 days	Survival, growth	[26]
Marine sediment chronic	Amphipod	10 days	Burrowing	[25]
Terrestrial	Earthworm (<i>Eisenia foetida</i>)	14 days	Survival	[27]
	Bobwhite quail	8 days	Survival, growth	[28]
	Mallard duck	8 days	Survival, growth	[28]

on individual organisms [12]. Reasons for doing a field study might include the fact that a current or future effect is likely, but there is not enough information to support a management decision, or it can provide additional evidence to link a site's contaminant and adverse effects. The key to conducting a good nonstandardized laboratory test or a field test is to find an appropriate reference site.

Studies of sites in which the contaminants of concern are found in the water should include the examination of the effects on periphyton, plankton, benthic macroinvertebrates, or fish [12]. Periphyton and plankton, such as algae, protozoa, and some crustaceans, are primary producers. Periphyton is typically collected by scraping or suction, or an artificial substrate may be deployed for a number of days to allow

attachment. Plankton is collected by pumping, netting, or trapping. Endpoints examined usually involve measures of community composition, such as species numbers, diversity, and richness.

Benthic macroinvertebrates are indicative of the effects of sediment contamination and are typically a good choice to examine if there are lipophilic contaminants at the site. They are collected by digging, netting, dredging, or grab sampling. Endpoints measured include the number and types of species, and other measures of species diversity. Macroinvertebrates can also easily be used in laboratory sediment toxicity tests, using either spiked artificial sediments or sediments taken from the site. Fish are good to examine because they can occupy different trophic levels, and they are relatively easy to collect by seining, passive netting, trawling, or electrofishing. Endpoints examined are often numbers and diversity, or sublethal effects such as behavior or tumor burden. Like benthic macroinvertebrates, fish are often used in laboratory toxicity tests.

Studies of sites in which the contaminants of concern are found in the soil often involve examination of effects on earthworms, nematodes, amphibians, mammals, or reptiles [29]. Earthworms and other soil nematodes are good for both metals and organics since they are in constant contact with the soil. Nematodes are collected by coring, sieving, or they can be driven from the soil by heat. Worms can be tested using spiked soil as the medium, or are examined directly on site, with typical endpoints being survival and reproduction [27]. Organisms such as amphibians, reptiles, and mammals are not used or collected as frequently. When they are, the risk assessor generally chooses a small species, since its home range is likely to be smaller and it tends to be more abundant than larger species. These types of organisms are generally collected by trapping, or their presence can simply be determined by examining surrogate measurements such as footprints, droppings, or carcasses [12]. Birds are not typically used or examined in a risk assessment because their home ranges can be quite large and their collection tends to be difficult and time intensive. When they are examined, they can be trapped to determine their numbers, or surrogate measures, such as auditory signals, used to determine their presence.

Alternatively, the number of nests can be counted and the success of each nest can be documented.

Other Assays Providing Support to Risk Assessments

Beyond using standard toxicity tests to assess survival, growth, and reproduction, or looking for the presence of an organism during a field survey, there are several different ways to determine other types of adverse impacts, such as altered community composition, to estimate body burdens of toxicants, and to examine markers or exposure and/or effects.

Biological Surveys

Two such methods to examine altered community composition are biotic indices and rapid bioassessment protocols. These two types of surveys compare population and community compositions of macroinvertebrates, fish, or periphyton between a reference site and the site of interest. For both of these protocols, an initial survey is conducted to quantify the presence and abundance of different species at both sites. These quantities are then converted into numerical values that relate to the overall health of the community.

More than 50 different indices have been developed to assess community health [10]. Parameters examined include species richness, or the number of species in a community, density, dominance, diversity (the abundance of individuals in one taxon versus all other individuals), as well as the number of sensitive species, such as the number of species belonging to the Ephemeroptera (mayflies), Plecoptera (stoneflies), and Trichoptera (caddisflies) or the EPT index [30]. Other metrics include percent dominant taxon, percent tolerant taxa, and percent Chironomidae, all of which are predicted to increase in response to increased perturbation [10]. The Family Biotic Index is also a good measure of perturbation, as it uses tolerance values to weight abundance in an estimate of overall pollution [10]. Other indices such as the Simpson's Diversity Index and the Shannon-Wiener Diversity Index are also used to account for the richness and the evenness of the different species that make up the total community [31]. Finally, determining the percentage of

individuals in each of the five functional feeding groups can be used to determine whether the trophic structure is balanced [32].

Since many of these indices require a lot of time, expertise, and money, rapid bioassessment protocols (RBPs) have been developed since the 1980s to obtain these same measurements in a more cost effective and efficient way. These protocols were developed from previously existing bioassessment methods that were used by many state agencies. The RBPs examine other parameters besides numbers of organisms. In the case of streams, these include the available cover, embeddedness, velocity and depth of the river or stream, sediment deposition, channel flow status, channel alteration, frequency of riffles, bank stability, vegetative protection, and riparian vegetative zone width [10]. RBPs are appropriate for distinguishing the existence and severity of impairment of the water body, as well as identifying the causes of impairment. They are also used to determine if a stream is supporting a designated aquatic life use specified in state Water Quality Standards, evaluating the effectiveness of controls and various restoration activities, and characterizing regional reference conditions. Additional information on both types of bioassessments and their uses is provided in another entry.

Biomarkers and Toxicokinetic Models

A second method to estimate exposure of an organism to a contaminant is to use a biomarker. Biomarkers are defined as a quantifiable change in a biochemical, physiological, or histological parameter that can be used as a measure of response to a xenobiotic. The idea behind biomarkers is that they can detect more subtle adverse effects than lethality at lower chemical concentrations, which would better protect the population or community. Biomarkers can be indicators of either exposure or effects, although exposure biomarkers are much more common. Examples of biomarkers routinely used in environmental toxicology studies include the induction of cytochrome P4501A (CYP1A) to indicate PAH exposure, acetylcholinesterase activity to indicate organophosphate or carbamate pesticide exposure, metallothionein induction to indicate exposure to several metals, vitellogenin to indicate exposure to

estrogenic compounds, and eggshell thinning to indicate organochlorine exposure. Additional information on biomarkers is provided in another entry.

A final strategy is to incorporate physiologically-based pharmacokinetic (PBPK) models into the risk assessment. These model the time course of disposition of toxicants in the whole organism, which gives an estimate of body burdens. In a PBPK model, the organism is represented by several different compartments in which toxicants might be expected to be absorbed in, distributed to, biotransformed in, and excreted from, such as the alimentary canal, blood, fat, liver, and kidney. Thus, one can use the model to predict concentrations of contaminants in an organism, rather than having to measure actual body burdens of contaminants, which can be time consuming and extremely expensive. Beyond cost, the models have other advantages, such as being able to provide a time course of distribution of toxicants to any organ or tissue, being able to estimate effects by changing physiological parameters such as food consumption, mobility, or reproductive status, and being able to use the same model to predict toxic action in different species. Although PBPK models can improve estimates in determining whether the contaminant will accumulate to high enough levels in the organism of interest to cause an adverse effect, there are a number of distinct disadvantages to their use compared to simply measuring body burdens. First and foremost, one needs information to build the model, such as blood volume, heart rate, breathing rate, and organ volumes. The fact that these parameters are often ill-defined for many species detracts from the power and utility of the model. Second, the mathematics, computational tools, and training needed to build and run the models may be out of reach for many risk assessors.

In spite of these obstacles, there are several PBPK models available for wildlife species, such as fish and polar bears, which can describe the body and tissue burdens of contaminants. For example, fish PBPK models have been built for rainbow trout and have been used to estimate tissue concentrations of paraoxon and cadmium [33–35]. Likewise, there is a PBPK model that estimates arsenic concentrations in the different organs of tilapia (*Oreochromis mossambicus*) [36]. Additionally, there are also models to assess ionizing radiation doses to a variety

of non-human species [37]. The models for polar bears are a bit more sophisticated than those built for fish, in that they can estimate tissue concentrations of organohalogenated compounds, such as dieldrin and PCBs, but can also be used to estimate risk of effects, such as adverse impacts on reproduction [38]. It is likely that they can be extended to other at-risk populations, such as orcas and the arctic fox [39].

There are a few models in reptiles and birds that were developed to be more applicable to veterinary uses, but could be used to estimate contaminant concentrations in wildlife. Examples include a model developed in red-eared sliders (*Trachemys scripta elegans*) to determine the pharmacokinetics and tissue distribution of buprenorphine, a potential analgesic drug for reptiles [40]; a model for midazolam developed in chickens, but able to accurately predict tissue residues in turkey, pheasant, and quail [41]; and a model used to determine tissue levels of lipophilic pesticides in chickens, ducks, geese, and turkeys [42]. Alternatively, there are many rodent models for environmental contaminants that could potentially be adapted for use in other species. One such recent example includes a PBPK model in rats for developmental exposure to polybrominated diphenyl ethers (PBDEs) [43].

Future Directions Using Animal Models in Risk Assessments

Some future directions for increasing the quality of risk assessment is to refine our understanding of sublethal effects rather than focusing on survival as an endpoint, and to better determine the mode of action of contaminants. One potential way to accomplish this goal is through the use of genomics – analyzing changes in gene, protein, and metabolite expression. Genomics has the advantage of examining thousands of genes or proteins simultaneously as opposed to traditional biomarkers, which often measure a single gene or protein at a time. Like biomarkers, genomic changes can be linked to both exposure and effect outcomes. For example, rather than looking for reproductive failure in a toxicity test, one could examine changes in genes involved in steroidogenesis. These genes should be differentially expressed at much lower toxicant concentrations, yet can still provide a predictive measure that

adverse reproductive effects are likely to occur [44]. Thus, genomic techniques are more sensitive than traditional toxicity testing or biomonitoring and can be used to detect early indicators of environmental perturbation.

One of the issues encountered when using genomic data in a risk assessment is that baselines or control patterns of expression need to be established for a variety of species. This is true for those organisms routinely used in risk assessments, such as fathead minnows and *Daphnia*, but even more so for nontraditional species and those organisms collected in field studies, since the variation in natural populations can be quite significant [45]. Second, the data generated from a genomic study cannot be easily translated into terms familiar to many risk assessors, such as a NOEL or EC₅₀ [46, 47]. Some recent studies have developed no observed transcriptional effect levels (NOTEL) for *Daphnia magna* exposed to copper, and these NOTELs were validated after exposing *Daphnia* to water samples collected upstream and downstream from two copper mines in California [48, 49]. However, there are almost no published studies using a NOTEL and relating that value to an adverse event. Third, there is often a lack of phenotype associated with the changes in mRNA levels, making it hard to interpret just what types of adverse effects would be occurring to organisms at the site [46, 47, 50]. Studies linking adverse effects, such as altered growth and reproduction, with transcriptomic changes have also been conducted (reviewed in [51]), but there are too few of these published to support the current use of genomics in risk assessments. Finally, genomic studies are expensive and the computational needs are quite high.

The current policy at EPA is that toxicogenomics data may be used in a weight-of-evidence approach for ecological risk assessment, but must be provided in conjunction with other information. At this point, genomics data alone are insufficient for a risk assessment and any subsequent management decision [52, 53]. However, some potential uses for genomics in risk assessment for the foreseeable future are to guide the experimental design of monitoring programs so as to reduce the number of animals used, as an initial screening to rank or set priorities for further toxicity testing, and to help extrapolate the effects of contaminants across species [54].

Bibliography

Primary Literature

1. EPA (1992) Framework for ecological risk assessment. EPA/630/R-92/001, Risk Assessment Forum, Washington
2. EPA (1998) Guidelines for Ecological Risk Assessment. EPA/630/R-95/002F. Risk Assessment Forum, Washington
3. Suter GW II (2007) Ecological Risk Assessment, 2nd edn. CRC, Boca Raton, p 643
4. Dale VH, Mulholland PJ, Olsen LM, Feminella JW, Maloney KO, White DC, Peacock A, Foster T (2004) Selecting a suite of ecological indicators for resource management. In: Kapustka L et al (eds) Landscape ecology and wildlife habitat evaluation: critical information for ecological risk assessment, land-use management activities, and biodiversity enhancement practices. ASTM International, West Conshohocken
5. Hill AB (1965) The environment and disease: association or causation? *Proc R Soc Med* 58:295–300
6. Fox GA (1991) Practical causal inference for ecoepidemiologists. *J Toxicol Environ Health* 33:359–373
7. EPA (2000) Stressor identification guidance document. EPA/822/B-00/025. Office of Water, Washington
8. EPA (1997) Environmental monitoring and assessment program (EMAP): research strategy. EPA/620/R-98/001. Office of Research and Development, Washington
9. EPA (2002) Research strategy: environmental monitoring and assessment program. EPA 620/R-02/002. Office of Research and Development, Research Triangle Park
10. Barbour MT, Gerritsen J, Snyder BD, and Stribling JB (1999) Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates, and fish, 2nd edn. EPA 841-B-99-002, ed. U.S.E.P.A.O.o. Water, Washington
11. Suter GW II (2006) Ecological risk assessment and ecological epidemiology for contaminated sites. *Hum Ecol Risk Assess* 12:31–38
12. EPA (1994) ECO update: field studies for ecological risk assessment. Publication 9345.0-051. Office of Emergency and Remedial Response, Washington
13. Smith PN, Cobb GP, Godard-Coddling C, Hoff D, McMurry ST, Rainwater TR, Reynolds KD (2007) Contaminant exposure in terrestrial vertebrates. *Environ Pollut* 150:41–64
14. EPA (1993) Wildlife exposure factors handbook. EPA/600/R-93/187. Office of Research and Development, Washington
15. EPA (1994) ECO update: using toxicity tests in ecological risk assessments. Publication 9345.0-051. Office of Solid Waste and Emergency Response, Washington
16. 04 AE (2004) Standard Guide for Selection of Resident Species as Test Organisms for Aquatic and Sediment Toxicity Tests. ASTM International, West Conshohocken, DOI: 10.1520/E1850-04
17. Landis WG, Yu M-H (2003) Introduction to environmental toxicology: impacts of chemicals upon ecological systems, 3rd edn. CRC, Boca Raton
18. 96 AE (2007) Standard guide for conducting acute toxicity tests on test materials with fishes, macroinvertebrates, and amphibians. ASTM International, West Conshohocken, DOI: 10.1520/E0729-96R07
19. 97 AE (2008) Standard guide for conducting acute toxicity tests on aqueous ambient samples and effluents with fishes, macroinvertebrates, and amphibians. ASTM International, West Conshohocken, DOI: 10.1520/E1192-97R08
20. 01 AE (2006) Standard guide for conducting three-brood, renewal toxicity tests with *Ceriodaphnia dubia*. ASTM International, West Conshohocken, DOI: 10.1520/E1295-01R06
21. E1193-97 AS (2004) Standard guide for conducting daphnia magna life-cycle toxicity tests. ASTM International, West Conshohocken, DOI: 10.1520/C0033-03
22. E1241-05 A (2005) Standard guide for conducting early life-stage toxicity tests with fishes. ASTM International, West Conshohocken, DOI: 10.1520/E1241-05
23. D3978-80 (1987) e1 (1997) Standard practice for algal growth potential testing with *Selenastrum capricornutum*. ASTM International, West Conshohocken
24. 98 AE (2004) Standard guide for conducting static acute toxicity tests starting with embryos of four species of saltwater bivalve molluscs. ASTM International, West Conshohocken, DOI: 10.1520/E0724-98R04
25. 03 AE (2008) Standard test method for measuring the toxicity of sediment-associated contaminants with estuarine and marine invertebrates. ASTM International, West Conshohocken, DOI: 10.1520/E1367-03R08
26. 05e1 AE (2005) Standard test method for measuring the toxicity of sediment-associated contaminants with freshwater invertebrates. ASTM International, West Conshohocken, DOI: 10.1520/E1706-05E01
27. 04 AE (2004) Standard guide for conducting laboratory soil toxicity or bioaccumulation tests with the lumbricid earthworm *Eisenia Fetida* and the Enchytraeid Potworm *Enchytraeus albidus*. ASTM International, West Conshohocken, DOI: 10.1520/E1676-04
28. 05e1 AE (2005) Standard practice for conducting subacute dietary toxicity tests with avian species. ASTM International, West Conshohocken, DOI: 10.1520/E0857-05E01
29. Golden NH, Rattner BA (2003) Ranking terrestrial vertebrate species for utility in biomonitoring and vulnerability to environmental contaminants. *Rev Environ Contam Toxicol* 176: 67–136
30. Wallace JB, Grubaugh JW, Whiles MR (1996) Biotic indices and stream ecosystem processes: results from an experimental study. *Ecol Appl* 6:140–151
31. Lampert W, Sommer U (1997) Limnoecology: the ecology of lakes and streams, 2nd edn. Oxford University Press, New York
32. Merritt RW, Cummins KW (1995) An introduction to the aquatic insects of North America, 3rd edn. Kendall/Hunt, Iowa
33. Abbas R, Hayton WL (1997) A physiologically based pharmacokinetic and pharmacodynamic model for paraoxon in rainbow trout. *Toxicol Appl Pharmacol* 145:192–201

34. Thomann RV, Shkrel F, Harrison S (1997) A pharmacokinetic model of cadmium in rainbow trout. *Environ Tox Chem* 16:2268–2274
35. Franco-Uria A, Otero-Muras I, Balsa-Canto E, Alonso AA, Roca E (2010) Generic parameterization for a pharmacokinetic model to predict Cd concentrations in several tissues of different fish species. *Chemosphere* 79:377–386
36. Liao CM, Liang HM, Chen BC, Singh S, Tsai JW, Chou YH, Lin WT (2005) Dynamical coupling of PBPK/PD and AUC-based toxicity models for arsenic in tilapia *Oreochromis mossambicus* from blackfoot disease area in Taiwan. *Environ Pollut* 135: 221–233
37. Beresford NA, Barnett CL, Brown JE, Cheng JJ, Copplestone D, Filistovic V, Hosseini A, Howard BJ, Jones SR, Kamboj S, Kryshev A, Nedveckaite T, Olyslaegers G, Saxén R, Sazykina T, Vives I, Battle J, Vives-Lynch S, Yankovich T, Yu C (2008) Inter-comparison of models to estimate radionuclide activity concentrations in non-human biota. *Radiat Environ Biophys* 47:491–514
38. Sonne C, Gustavson K, Rigét FF, Dietz R, Birkved M, Letcher RJ, Bossi R, Vorkamp K, Born EW, Petersen G (2009) Reproductive performance in East Greenland polar bears (*Ursus maritimus*) may be affected by organohalogen contaminants as shown by physiologically-based pharmacokinetic (PBPK) modelling. *Chemosphere* 77:1558–1568
39. Sonne C (2010) Health effects from long-range transported contaminants in Arctic top predators: An integrated review based on studies of polar bears and relevant model species. *Environ Int* 36:461–491
40. Kummrow MS, Tseng F, Hesse L, Court M (2008) Pharmacokinetics of buprenorphine after single-dose subcutaneous administration in red-eared sliders (*Trachemys scripta elegans*). *J Zoo Wildl Med* 39:590–595
41. Cortright KA, Wetzlich SE, Craigmill AL (2009) A PBPK model for midazolam in four avian species. *J Vet Pharmacol Ther* 32:552–565
42. MacLachlan DJ (2010) Physiologically based pharmacokinetic (PBPK) model for residues of lipophilic pesticides in poultry. *Food Addit Contam A Chem Anal Control Expo Risk Assess* 27:302–314
43. Emond C, Raymer JH, Studabaker WB, Garner CE, Birnbaum LS (2010) A physiologically based pharmacokinetic model for developmental exposure to BDE-47 in rats. *Toxicol Appl Pharmacol* 242:290–298
44. Kennedy SW, Euling S, Huggett DB, De Coen LW, Snape JR, Zacharewski TR, Kanno J (2008) Application of genomics to screening. In: Ankley GT et al (eds) *Genomics in regulatory ecotoxicology*. CRC, Pensacola
45. Oleksiak MF, Churchill GA, Crawford DL (2002) Variation in gene expression within and among natural populations. *Nature Genet* 31:261–266
46. Ankley GT, Daston GP, Degitz SJ, Denslow ND, Hoke RA, Kennedy SW, Miracle AL, Perkins EJ, Snape J, Tillitt DE, Tyler CR, Versteeg D (2006) Toxicogenomics in regulatory ecotoxicology. *Environ Sci Technol* 40:4055–4065
47. Boverhof DR, Zacharewski TR (2006) Toxicogenomics in risk assessment: applications and needs. *Toxicol Sci* 89: 352–360
48. Poynton HC, Loguinov AV, Varshavsky JR, Chan S, Perkins EJ, Vulpe CD (2008) Gene expression profiling in *Daphnia magna* part I: concentration-dependent profiles provide support for the No Observed Transcriptional Effect Level. *Environ Sci Technol* 42:6250–6256
49. Poynton HC, Zuzow R, Loguinov AV, Perkins EJ, Vulpe CD (2008) Gene expression profiling in *Daphnia magna*, Part II: validation of a copper specific gene expression signature with effluent from two copper mines in California. *Environ Sci Technol* 42:6257–6263
50. Degitz SJ, Hoke RA, Bradbury S, Brennan R, Ferguson L, Klaper R, Orban L, Spurgeon D, Tilton S (2008) Application of genomics to regulatory ecological risk assessments for pesticides. In: Ankley GT et al (eds) *Genomics in regulatory ecotoxicology*. CRC, Pensacola
51. Van Aggelen G, Ankley GT, Baldwin WS, Bearden DW, Benson WH, Chipman JK, Collette TW, Craft JA, Denslow ND, Embry MR, Falciani F, George SG, Helbing CC, Hoekstra PF, Iguchi T, Kagami Y, Katsiadaki I, Kille P, Liu L, Lord PG, McIntyre T, O'Neill A, Osachoff H, Perkins EJ, Santos EM, Skirrow RC, Snape JR, Tyler CR, Versteeg D, Viant MR, Volz DC, Williams TD, Yu L (2010) Integrating omic technologies into aquatic ecological risk assessment and environmental monitoring: hurdles, achievements, and future outlook. *Environ Health Perspect* 118:1–5
52. Dix DJ, Gallagher K, Benson WH, Groskinsky BL, McClintock JT, Dearfield KL, Farland WH (2006) A framework for the use of genomics data at the EPA. *Nat Biotechnol* 24:1108–1111
53. EPA (2007) Interim guidance for microarray-based assays: data submission, quality, analysis, management, and training considerations. Draft, Science Policy Council, Washington
54. Perkins EJ, Denslow N, Chipman JK, Guiney PD, Oris JT, Poynton H, Robidoux PY, Scroggins R, Van Der Kraak GJ (2008) Application of genomics to assessment of the ecological risk of complex mixtures. In: Ankley GT et al (eds) *Genomics in regulatory ecotoxicology*. CRC, Boca Raton

Books and Reviews

- Barnthouse LW, Munns JWR, Sorensen MT (eds) (2008) *Population-level ecological risk assessment*. CRC, Boca Raton
- Kapustka L, Biddinger G, Luxon M, Galbraith H (eds) (2004) *Land-scape ecology and wildlife habitat evaluation: critical information for ecological risk assessment, land-use management activities, and biodiversity enhancement practices*. ASTM International, West Conshohocken
- Solomon KR, Brock TCM, de Zwart D, Dyer SD, Posthuma L, Richards SM, Sanderson H, Sibley PK, van den Brink PJ (eds) (2008) *Extrapolation practice for ecotoxicological effect characterization of chemicals*. CRC, Boca Raton
- Ware GW (2006) *Reviews of environmental contamination and toxicology*, vol 187. Springer, New York

Ecological Succession and Community Dynamics

HERMAN H. (HANK) SHUGART

Department of Environmental Sciences,
The University of Virginia, Charlottesville, VA, USA

Article Outline

Glossary
Definition of the Subject and Its Importance
Introduction
Sustainability and Vegetation Dynamics
Future Directions
Bibliography

Glossary

Abstract community A group of organisms that recurs on the landscape with an implication of a level of integration among its parts that in extreme could be called organismal or quasi-organismal (see: Concrete Community).

Chronosequence A space-for-time substitution in which the successional vegetation is ordered in a regular fashion. An example would be the series of vegetation at the foot of a receding glacier or a series of sand dunes ordered in regularly aged series.

Clementsian succession An explanation of succession emphasizing the attributes of the community as if it functioned like a single living organism (see: Abstract Community).

Climax community An ecological community associated with a particular climate and in a state of dynamic equilibrium with the climate. The American ecologist, F.E. Clements, designated the climax community as the endpoint of ecological succession for a given climate condition and noted the common Greek root *κλίμα* (*clima*) or inclination, in both climax and climate as indicating their close relationship.

Concrete community The collection of organisms found at a specific place and time (see: Abstract Community).

Gleasonian succession theory An explanation of succession emphasizing the importance of the attributes of individual organisms as the fundamental basis.

Indicator species Plants or animals whose presence implies the past or present conditions at a given location.

Monoclimax The concept that one mature vegetation will eventually be produced by successional processes in a given region.

Polyclimax The concept that multiple mature, stable-vegetation types can develop from the successional processes operating in a given region.

SIGMATIST A school of vegetation science emphasizing the classification of vegetation.

Time-for-space substitution The collection of vegetation data from different locations at which succession has been initiated at different times in the past to piece together the pattern of succession.

Definition of the Subject and Its Importance

“Ecological Succession” is an ordered progression of structural and compositional changes in communities toward an eventual unchanging condition, the climax community [1–3]. The term “Community” is used in two ways [4]. The “Abstract Community” refers to an abstract group of organisms that recurs on the landscape, a definition, which usually carries with it an implication of a level of integration among its parts that in extreme could be called organismal or quasi-organismal; the “Concrete Community” concept refers to the collection of organisms found at a specific place and time. These terms and their meanings are topics of significant debate among ecologists, both historically and today [5, 6]. These differences in the meaning and cause of ecological succession strongly affect the formulation of policies for ecosystems management and restoration.

Succession is important in sustainability science for several reasons: Many of the laws and regulations regarding the maintenance and restoration of terrestrial ecosystems contain assumptions that are products of different theories as to how ecological succession proceeds in a given environment. Laws might require restoration to a natural condition, which is often the

vegetation that is considered the stable single vegetation type on a map that is itself based on the idea that vegetation undergoes succession a vegetation type appropriate to the climate (the climax vegetation). Further, the management of land systems often has as their basis concepts as to how succession works. For example, if one subscribes to a successional theory that abused land under an inappropriate management scheme will return to its natural state if the inappropriate management is relaxed, then one can utilize land until monitoring indicates the ecosystem is over stressed. If an alternate successional theory is applied, that abused land may not recover to its former when released from inappropriate management, then a much more cautious form of management is needed.

Introduction

Ecological succession and the dynamic change of communities is one of the pillars of modern ecology and it has considerable influence on sustainability science. The processes that drive succession and the regularly of patterns of succession have been the topic of considerable debate over the past century – due in no small part to the importance of the concept to basic and applied ecology. Succession is at its simplest level the pattern of change in ecological communities and these changes can be read on the landscape if one knows regional history or conversely can be used to interpret a landscape's history. In the former case, the historical pattern of land abandonment in the piedmont of the southeastern United States presents land elements abandoned and never reclaimed since the US Civil war in 1860, land left fallow in the First World War; farms left to ruin in the crushing economic Great Depression of the 1930s; more land left when the sons of farmers left for the Second World War; still other farm land taken out of production in agricultural programs reduce farm acreage in the 1950s. The landscape as one drives and automobile through Georgia or the Carolinas, is a kaleidoscopic montage with similar parcels of land cover types replicated on locations with similar history. In the latter case, a small patch of pine (*Pinus* sp.) trees on a slope bears witness to recovery from a past wildfire or a Beech (*Fagus*) forest of great stature indicates an area that has not burned, perhaps

for centuries. The dependence of pine species on fire for regeneration or the sensitivity of the thin-barked beech to fire is written into the landscape response.

At the turn of the twentieth century, an American plant ecologist, F.E. Clements, framed many of the definitions and concepts in ecological succession and produced a theoretical system of concepts that explained the nature and function of ecological communities changing over time. Clements' definitions and approaches to understanding the dynamics of landscapes were used to develop regional systems of land management policies based on expectations, succession, and successional recovery. Clements also developed a niche-based concept of using what he called "indicator species" to evaluate the history and condition of land units [2]. For example, an overabundance of bitter or poisonous herbs might be indicators of overgrazing in a prairie systems; the occurrence of other species might indicate unusual soil chemistry and thus the presence of valuable minerals in the bedrock. The details of Clements' ideas were debated strongly by his contemporaries and other ecologists in the 1920s and 1930s found differences with parts of his successional concepts. For example, H.A. Gleason [7, 8] argued that the highly coordinated interactions in Clements concepts had at their basis the unique attributes of the individual plant species. These will be discussed below. Perhaps the most sharply pointed criticism was from the British ecologist, A.G. Tansley, who argued that the Clements' view of succession had so clouded the meaning of the term, community, that a new term, the ecosystem [5], was needed to replace and properly formalize these ideas. It is significant in understanding the emphases in ecosystem studies to note that A.G. Tansley was a coauthor to the 1899 edition of Spencer's *Principles of Biology*, the leading English biology text of the late nineteenth century. Tansley's definition of the ecosystem and his lack of acceptance of Clements' and Phillips' holistic view that the ecological community was a coherent, organism-like entity was not a rejection of a holistic view of ecology, it was a debate among holistic ecologists [9]. These and other continuing and sometimes rancorous [6, 10] debates will be discussed in more detail below.

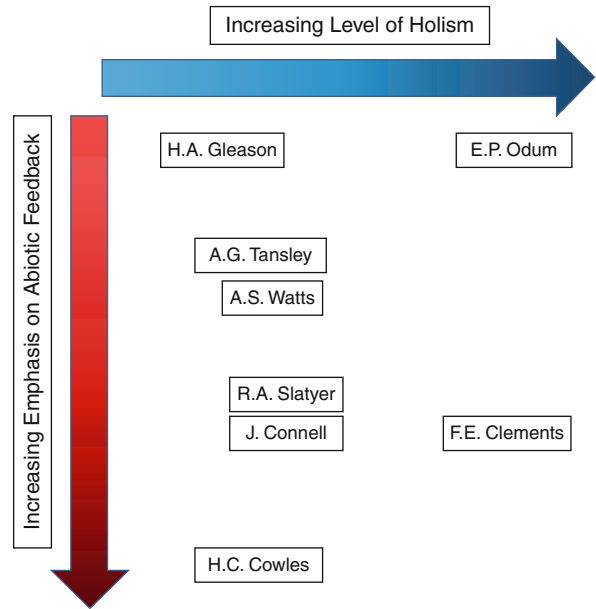
The development of the attributes of ecological succession has from its beginning had two poles: one

Clementsian pole, which emphasizes holism, great interconnectedness of ecological systems and a successional progression toward an ultimate stable ecological system with an optimization of its attributes; a second Gleasonian pole, which sees the properties of ecological systems as deriving from the individual attributes of the organisms that comprise it and their interactions with their environment. The Gleasonian view is less dogmatic about the necessary outcomes of succession. Of course, sustainability science ultimately must take the measure of the dynamics of complex ecosystems, the natural and direction of their change under management. Debates among ecologists as to the nature of succession must be understood by scientists seeking to understand sustainability.

It is important to note that through the history of this strong Anglo-American debate on the nature of succession, there has been a continental European effort toward the goal of developing a taxonomic tradition in the issue of classifying vegetation into meaningful units, which is also an essential element of sustainable management. The so-called Zurich-Montpellier school (or as later called the SIGMATIST school) originated in the 1890s with work by C. Schröter and Charles Flahault. This effort was taken up and lead by J. Braun-Blanquet. This tradition emphasized a floristic description of vegetation and had the ultimate goal a classification of “associations” as the basic vegetation units of interest [11]. This interest in vegetation classification as a basis of mapping and prediction continues to the present. Significantly, vegetation classification is an ultimate element in organizing management plans and hence for sustainability as well.

Historical Development of the Concept of Ecological Succession

Different scientists in different locations have observed these patterns and drawn rather different conclusions about the root causes of succession and to the degree that it repeats itself over time and space. Figure 1 illustrates the relationships among several ecologists who have developed what might be best described as “iconic” views of succession. The two axes in the figure are the level of holism in the theories and the emphasis on the degree of importance of feedbacks between the biological community and the environment.



Ecological Succession and Community Dynamics.

Figure 1

Iconic ecologists and their emphases on the attributes of successional systems

All of these scientists are holistic in that they focused on the larger system behavior of ecological systems. The most holistic of these scientists would see the community as a highly organized, perhaps highly coevolved systems described as an abstract community in the definition section of this paper (above). They also vary on the degree to which they emphasize feedbacks between the organisms and their environment.

F.E. Clements The iconic scientists in Fig. 1 represent different points of view in the formative debates on the nature of ecological succession in the formative period in the first third of the twentieth century. Clements was in some sense the dominant figure in these discussions. A.G. Tansley who represented strong opposition to many of Clements’ fundamental concepts, wrote Clements’ obituary [12] for *The Journal of Ecology*,

- ▶ Though out of accord with many American ecologists, Clements had, throughout his career, a devoted following of younger men, and besides exercising a worldwide influence through his theory of vegetation, he

directly inspired a great deal of American ecological research. This was partly due to his powerful personality. He was decidedly puritan, even ascetic (he neither drank nor smoked, and it gave him real pain to see other people doing so), and his manner was apt to be tinged with a certain arrogance. These things naturally antagonized many people; but at the same time his capacity for hard work, his intense and complete devotion to his subject, his powerful intellect and unremitting search for fresh knowledge and satisfactory formulations, could not fail to inspire the highest respect.

Clements' concept [3] saw secondary succession as being initiated by an event that removed all of the vegetation. Nowadays, this would fall under the term "disturbance" [13]. Clements referred to this process as *nudation*. Seeds, spores, etc., would then disperse to the site in the process of *migration*. These would germinate and become established in a collective set of processes that he called "ecesis." The resultant plants would then interact with one another, notably through competition. Clements called these collection of interactions, "coaction." Successful plants would interact with the microenvironment at the site and, in a process called "reaction," would modify the local site conditions as to make it unsuitable for the plant species that were there and more suitable for other species, which would replace them. This interaction is often called "facilitation" in modern ecology texts. The plant communities generated by this process are called "seres." The feedback between each sere and the environment produces an ordered sequence of seres or seral stages of succession. This reaction-replacement process would continue until an assemblage of species, which were able to occupy the site after the plant-environment reaction process, became dominant ("stabilization"). This ultimate assemblage was the "climax community."

Clements' climax community concept is incorporated almost subliminally in a number of the ideas in plant ecology and landscape management, particularly in the Western United States. Maps of "potential vegetation" are often maps of climax vegetation in the style of Clements. This is reinforced by the two-dimensional nature of paper making it cumbersome to print maps with more than one potential vegetation in a given map location. Management of parks toward

a "natural" condition often has a "climax community" concept as an underlying basis. Decisions to sustainably conserve land are straight-forward if there were a climax vegetation but should be difficult for successional seres, which might be expected to change over time. While Clements' concepts are ubiquitous in land management, several ecologists have noted difficulties with the concept. More frequent criticisms of Clements' concepts include:

1. The idea that the sere/environment feedbacks produce a necessary ordering of communities toward the climax community. Clements [11] noted, "The climax formation is the adult organism of which all the initial stages are but stages of development." Some of the more ardent Clementsian ecologists [14–16] likened the sequence of seres in succession as a direct analog to the embryological development of an organism. The rejection of this concept of the community as a superorganism concept by Tansley [5] motivated the invention of the neologism, "ecosystem," as a replacement for what he felt was a corrupted term, "community," in 1935.
2. The concept of facilitation. Facilitation has been attacked from evolutionary arguments to the effect that one would not expect a species to evolve to facilitate the success of a replacing competitive species. Field observations of species replacements support the existence of facilitation such as the presence of plants with nitrogen-fixing symbiotic organisms increase the nutrient status of a site and advantaging other species. There are also numerous counterexamples of species that hold sites tenaciously over generations before finally losing to competitors [6].
3. The progressive ecesis of different species relayed to the site for all the seral stages. In some successional sequences, all the species are present at the site from the first nudation event. In these cases, the progression of seral stages is a consequence of the differences in rate of maturation of herbaceous plants, shrubs, and trees [10].
4. The existence of a single stable climax community reached by a single sequence of seres. Several questions have arisen about the climax concept. Is there a single climax community or is there always change in the vegetation at a site? [17] Can there be multiple stable climaxes? [18–20] Can there be several

pathways to a particular community? [21] Can the pathways fork? Are the seres comparable at different locations?

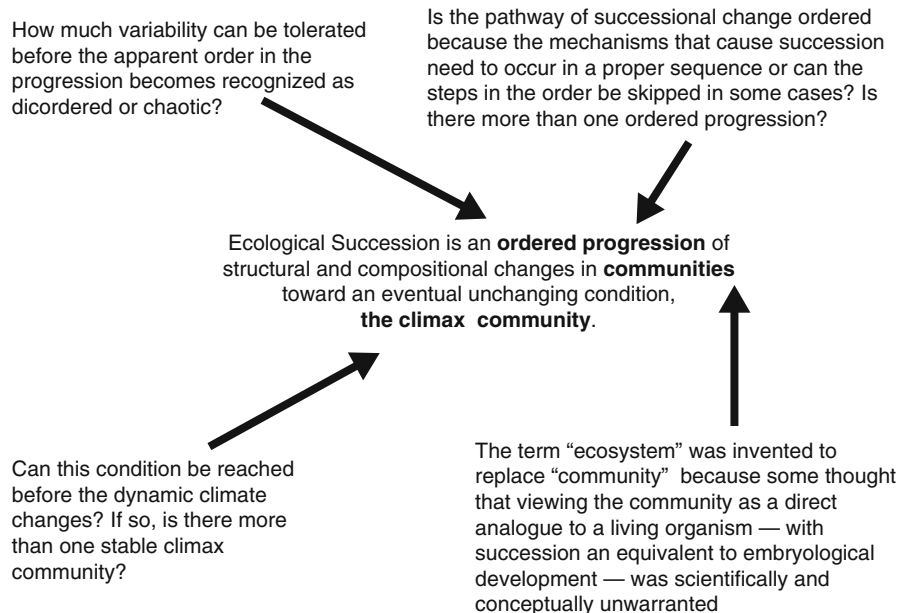
These criticisms continue in no small part because of the importance of Clements and his students in the formulation of a grand theory of vegetation dynamics and their ability to translate their concepts into practical applications. Subsequent recognition that past conditions could leave stable relict communities (e.g., preclimaxes such as grassland patches in temperate deciduous forest). Human or natural disturbances (such as occasional wildfires) could leave in a region a persistent vegetation that could be expected to change to a climax vegetation if the disturbance were controlled (“disclimax”). Unusual soil conditions or variations in topography could produce persistent vegetation that was different from the regional stable vegetation (“preclimaxes and post climaxes”). These elaborations of the climax concept inspired a proliferation of conditional climax vegetations [3]. These extended definitions of the Clementsian succession concept that essentially preserved the ideas of a single succession to a climax community through an ordered series of seral communities, also allowed for

special conditions (disturbance regime, human use of the land, historical antecedents and soils) to create deviations from this basic model [22].

The debate over these issues and its continuation over almost a century in some cases, speaks to the complexity of the issues and the importance of understanding succession in designing concepts to manage natural systems sustainably [11, 23]. The issues raised by the scientists in Fig. 1 are illustrated in Fig. 2 as deconstructions of the standard definition of ecological succession.

H.A. Gleason and the Individualistic Concept of Succession

Henry Gleason represents a logical foil to Clements holistic idea of progression of succession toward a climax vegetation. In total, his publications were strongly focused on plant taxonomy [24] but the relatively smaller section of ecological papers that he produced had remarkable impact on his colleagues, an impact that continues to the present day. As early as 1910, Gleason [25] noted, “. . . it is impossible to state whether there is one definite climax vegetation in each province; it seems probable that there are several such association each characteristic of a limited proportion.”



Ecological Succession and Community Dynamics. Figure 2

Some of the basic issues in the nature of ecological succession arising from a standard definition

Gleason's view of succession differed significantly from that of Clements. He recognized as early as 1908 [26] that succession to be retrograde as well as progressive as posited by Clements. Succession was not necessarily an irreversible trend toward the climax community. Gleason also recognized that the climate could change in contrast with Clements' theory of vegetation undergoing succession in an unchanging climate. Gleason developed what he called the "individualist concept" that succession is the result of environmental requirements of the individual species that comprise the vegetation. He also noted [27] that "... no two species make identical environmental demands."

Gleason's concept of succession was a much more fluid and much less stereotypical concept of succession than that of Clements. Succession reflected the interactions of individuals with the environment. It could change in its nature with different climatic and other environmental conditions. It could progress or regress to a more different, stable community depending on time and circumstance.

Henry Chandler Cowles One of the founders of ecological concepts of plant communities in the United States was the geologist turned ecologist, Henry C. Cowles, at the University of Chicago [11]. Cowles [28] studied the pattern of ecological succession in a set of sand dunes, the "Indiana Dunes," which are now protected by the State of Indiana's "Indiana Dunes State Park." These dunes were formed sequentially from sediment that was exposed with variation in Lake Michigan. The oldest of the dunes, which is the furthest away from the Lake Michigan beach, has a loamy brown forest soil and is covered by a Beech-Maple (*Fagus-Acer*) forest; the youngest dune at the beach front has windblown sand as a soil and is covered with a patchy beach-grassland. Cowles recognized this entire land pattern represented a chronosequence, with the spatial location of the dunes of different ages representing different stages of successional development. He later [29] produced the now famous phrase that "succession was a variable chasing a variable" – the changes in vegetation chase the changes in the climate. This insight was, and remains, a remarkable concept of vegetation function. What he meant by this was that the rate of change of vegetation succession, the variable, is sufficiently slow that one could expect the climate,

the chased variable, to change by the time the succession process was completed.

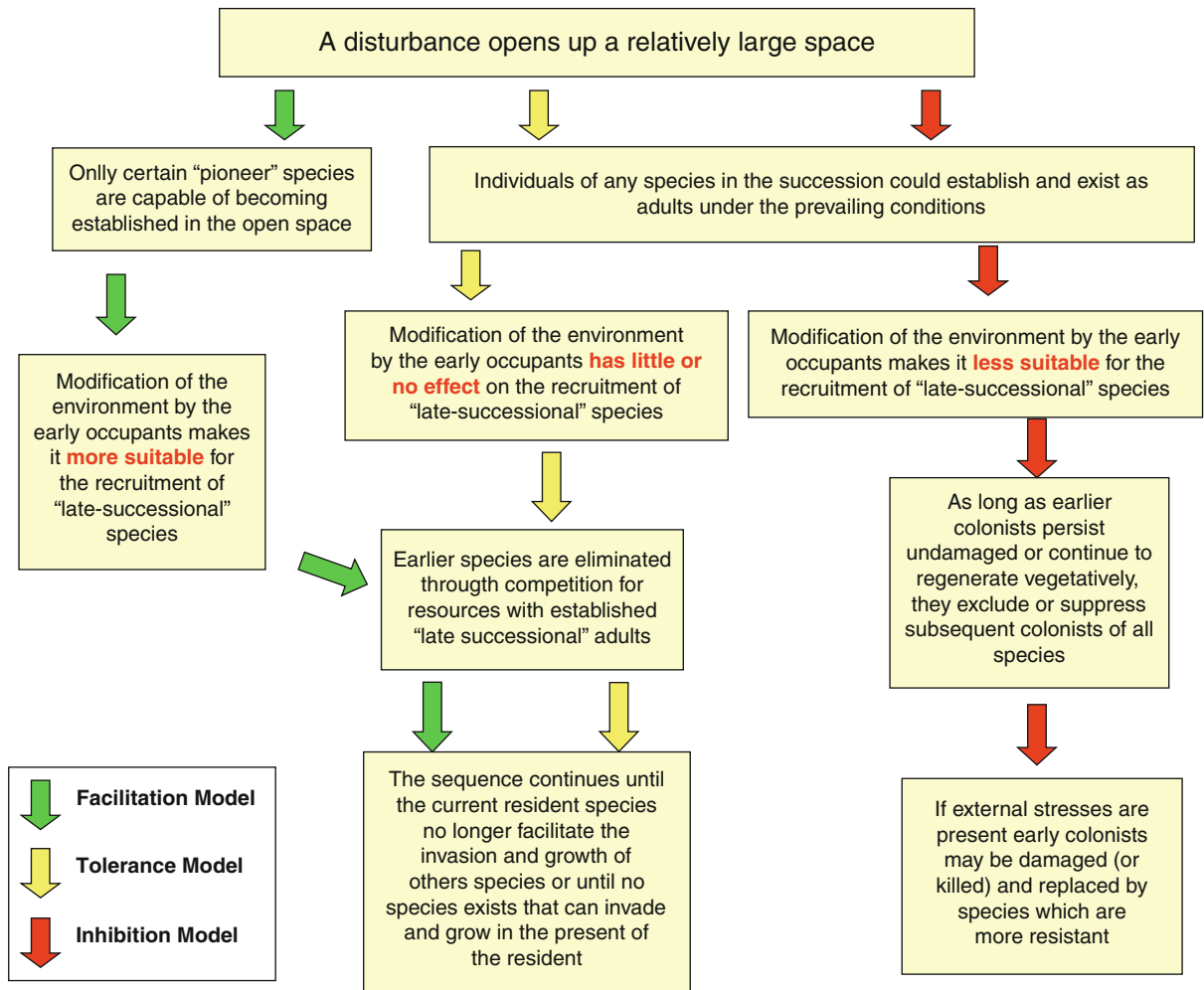
Cowles was a remarkable and innovative scientist. He was honored in the 1935 issue of *Ecology*, the journal of the Ecological Society of America. The Henry Chandler Cowles issue (*Ecology*, Volume 16, Number 3) is a collage of the central issues in plant ecology in its formative years: C.C. Adams, [30] an early researcher of ecological succession and the associated bird communities, along with F.E. Clements, [31] wrote on human ecology and ecology in the public service; Transeau [32] discussed the "Prairie Peninsula," a region of prairie vegetation jutting into a forests in the American Midwest, and emphasized that one must consider past climates as well as present climates to understand vegetation; Fuller [33] and Sears [34] made much the same point looking at paleoecological data; several students and colleagues of Cowles presented data on mature vegetation in different areas.

A.G. Tansley and the Ecosystem Concept The first of the papers in the Cowles issue and different in tone from the rest, was Tansley's now classic 1935 paper on "The use and abuse of vegetational concepts and terms." Tansley [5] discussed the ideas of the American, F.E. Clements, and the amplification of these ideas by the South African, John Phillips. Tansley declared his strong disagreement with three of Phillips' papers in the *Journal of Ecology* [14–16] characterizing ecological communities as "quasi-organisms" whose successional dynamics were analogous to embryological development, "The community is born, grows, matures, reproduces, and carries out various other biotic phenomena: it behaves in a manner similar to an individual, with obvious and natural differences inherent in its wholly divergent and far more complex structure, constitution and functions. It behaves in such a way as to justify the view that not only is it similar to an organism, but that it is a kind of organism . . ." This conceptualization is in keeping with the principle of holism from Aristotle's *Metaphysics* onward and often abbreviated as "the whole is greater than the sum of its parts." Phillips attributed many of his ideas to Jan Smuts' book, *Holism and Evolution* [35] and noted later in writing a eulogy for Clements, [36] that his colleague Smuts was profoundly influenced by the writings of Clements.

Tansley created the ecosystem concept as in contrast to Phillips' and Clements' views. It is significant that in his iconoclastic "ecosystem" definition, he emphasized ecosystems "were of the most various kinds and sizes." In doing so, Tansley basically defined an ecosystem as what systems scientists would nowadays call a *system of definition*: an arbitrary system defined by the specific considerations for a particular application [37]. Tansley's ecosystem definition conforms well to more mathematical, interactive-system concepts in other sciences. His definition includes the intrinsic

consideration of scale found in other sciences, particularly physically based sciences.

Tansley also endorsed the concept of the "polyclimax", the possibility of multiple stable ecosystems following from succession and a logical contrast to the Clements idea of the "monoclimax" as the eventual product of successional processes working in a region. An example of the polyclimax is shown in Fig. 3 for succession after fires in the boreal forests in the vicinity of Fairbanks, Alaska. In this case, the pattern of succession following a wildfire on



Ecological Succession and Community Dynamics. Figure 3

Three different models of ecological succession from J.H. Connell and R.O. Slatyer [19] depicting succession as arising from the resident species at a location encouraging the colonization of a subsequent set of species (Facilitation Model), by the resident species being relatively neutral to the other species (Tolerance Model), or by the resident species inhibiting the colonization by other species (Inhibition Model)

south-facing slopes eventually leads to a White Spruce (*Picea glauca*) forest that is quite different in its composition and character from the Black Spruce (*Picea mariana*) forest seen on north-facing slopes. To some degree, Clements and his colleagues used terms such as disclimax, etc. (see above) to handle these sorts of cases in their monoclimate theory of succession.

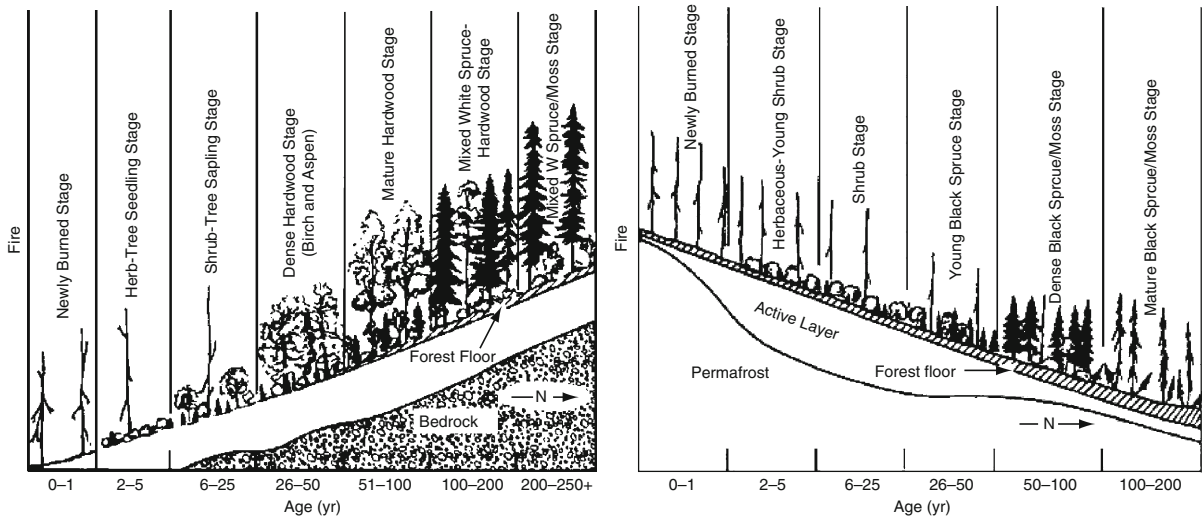
A.S. Watt One of Tansley's outstanding students was Alexander Stuart Watt. He developed his doctorate in 1924 on the dynamics of Beech (*Fagus* sp.) forests [38]. He noted the structure of the beech forest was at the scale of the canopy controlled by a dominant large tree best characterized as a cyclical process of the regeneration in the patch of well-lighted area that followed the death of a canopy tree, the growth of these recruits and their competition with one, and the development of a dominant canopy tree in that location. The death of this large tree would then reinitiate the cycle. By 1947, Watt had become the president of the British Ecological Society and delivered one of the most important papers in ecology as his presidential address. This presidential address and the subsequent paper represented a synthesis of what nowadays might be called the "big picture" science view of his doctoral advisor, A.G. Tansley, with Watt's own insights based of his field experience and broad ecological interests. In his address, Watt represented the plant community as a "working mechanism" of interacting plant processes such as regeneration, growth, and death on individual plants producing the broader-scale pattern of plant communities. He stated, "The ultimate parts of the community are the individual plants, but a description of it in terms of the characters of these units and their spatial relations to each other is impractical at the individual level. It is, however feasible in terms of the aggregates of individuals and species which form different kinds of patches: these patches form a mosaic and together constitute the community. Recognition of the patch is fundamental to an understanding of the structure as analyzed here" [39]. To illustrate this communality in the underlying working dynamics of vegetation, he compared a diverse array of plant communities – heathlands, grasslands, bogs, montane vegetation and forests.

As was the case with Gleason, Watt noted that one must account for the results of individual plants

interacting with one another, but he also felt such an accounting was "impractical at the individual level." Twenty years later, the explosive expansion of computational power that still continues to date has made such computations more and more feasible. With the development of increasingly powerful digital computers, starting in the 1960s and continuing to the present, several different scientific disciplines (physics, astronomy, ecology) independently began to apply computers to the tasks of "book-keeping" the changes and interactions of individual entities. Forest "gap" models (which will be discussed below) were one of this class of individual-based models or IBM's that brought Watt's insights on vegetation dynamics into a form that could be projected using computer simulation techniques.

Attempts at Synthesis: Connell and Slatyer The persistent dichotomy represented by Clementsian versus Gleasonian views of succession has consistently invited attempts to synthesis toward a unified concept. J.H. Connell and R.O. Slatyer [19] developed a synthesis that emphasized the individual attributes of plants and their feedback with their environment. Clementsian succession was a special case, which they called the "Facilitation Model" (Fig. 4). The replacement patterns in the Connell and Slatyer model were for sets of species, which also implies to some degree a replacement of one community with another, which also is more of a Clementsian concept. Succession was driven by transfers from one set of species as the facilitation model in which succession arises from the resident species at a location encouraging the colonization of a subsequent set of species, as the tolerance model in which the resident set of species are relatively neutral in their interaction with other species, or as the inhibition model in which the resident species block the colonization by other species (Fig. 4).

Attempts at Synthesis: E.P. Odum Eugene Pleasants Odum was a significant figure in what Robert McIntosh called the rise of ecosystems science. In 1953, he wrote the first modern ecology textbook, [40] which was the standard text for an emerging formulation of ecology as a formal academic discipline. He and his brother H.T. Odum were instrumental in the use of energy transfers in ecosystems (ecological energetics) as



Ecological Succession and Community Dynamics. Figure 4

Ecological succession following wildfires in the vicinity of Fairbanks, Alaska (USA). *Left*. Ecological succession of a south facing slope. *Right*. Ecological succession on a north-facing slope [78]

a common currency for comparing ecosystems with one another. As the sons of a crusading sociologist, H.W. Odum, the two Odum brothers were drawn to include “human ecosystems” in their studies, particularly the energetics of human society and its interactions with the natural ecosystems. The impact of ecological energetics on ecology was profound in that it promoted large-system intercomparisons. The International Biological Programme, the first “big picture” synoptic study of ecosystems used ecological energetics as an organizing principle for projects that compared forests, grasslands, and other biomes, worldwide. These comparisons were often made using “compartment models,” diagrams in which the transfers of energy in units of such as $\text{Kcalm}^2 \text{ year}^{-1}$ were indicated by arrows connecting boxes which expressed the integral of the arrows such as Kcalm^2 . Compartment models were easily transformed into ordinary differential equations for model projections

Also in the 1950s and 1960s, the United States Atomic Energy Commission (USAEC) sponsored a large program studying the movement of radioactive isotopes in natural environments including whole ecosystem experiments and direct measurements of rates of transfer of isotopes through natural environments [41]. Large, linear compartment models developed in pharmacology [42], and used in ecological energetics

were developed to predict the ecological transfer of these materials. There were several advantages to such linear models. They (even for complex linear models) can be solved at equilibrium by algebraic manipulations, an advantage given the modest computational power available in the 1950s and 1960s. When the transfer of material between compartments is a fixed percentage of the amount of material in a source compartment (called donor-controlled flow), linear models have the property that, with a constant input of material into the system, the equilibrium levels of materials are also the maximum amount. Importantly for models of radioisotope transport, radioactive decay of the isotope is a linear, donor-controlled process. Models of this formulation continue today in applications to determine the transfer of radioactive material to human populations from reactors or nuclear accidents such as the Chernobyl event.

In 1969, E.P. Odum wrote “The Strategy of Ecosystem Development,” [43] a synthetic paper strongly based on ecological energetics and compartment models in ecosystems. The emphasis of the paper was strongly in the area that nowadays might be termed as sustainability science as is evidenced by its first lines, “The principles of ecological succession bear importantly on the relationship between man and nature. The framework of successional theory

needs to be examined as a basis for resolving man's current environmental crisis." Odum's view of succession was summarized as a tabular model of ecological succession (Table 1). It was strongly Clementsian with

Ecological Succession and Community Dynamics.

Table 1 A "tabular model" of changes expected over ecological succession from E.P. Odum

Ecosystem attributes	Early succession	Late succession
<i>GPP/respiration</i>	> or <1	Approaches 1
<i>GPP/biomass</i>	High	Low
<i>Biomass/energy</i>	Low	High
<i>Net community production (NPP)</i>	High	Low
<i>Food chains</i>	Linear	Web-like
<i>Total organic matter</i>	Small	Large
<i>Inorganic Nutrients</i>	Extrabiotic	Intrabiotic
<i>Species richness</i>	Low	High
<i>Species evenness</i>	Low	High
<i>Biochemical diversity</i>	Low	High
<i>Stratification and pattern</i>	Poorly organized	Well organized
<i>Niche specialization</i>	Broad	Narrow
<i>Size</i>	Small	Large
<i>Life cycles</i>	Short, simple	Long, complex
<i>Mineral cycles</i>	Open	Closed
<i>Nutrient exchange</i>	Rapid	Slow
<i>Role of detritus</i>	Unimportant	Important
<i>Selection on growth form</i>	r-Selection	k-Selection
<i>Selection on production</i>	Quantity	Quality
<i>Symbiosis</i>	Undeveloped	Developed
<i>Nutrient conservation</i>	Poor	Good
<i>Stability</i>	Low	High
<i>Entropy</i>	High	Low
<i>Information content</i>	Low	High

GPP gross primary production (photosynthesis), *NPP* net primary production (GPP minus respiration)

Source: From Odum [40]

some of the superorganism emphasis of Phillips added in. It also included a number of ideas that were popular in theoretical and population ecology at the time, such as the idea that biotic diversity in an ecosystem connotes ecosystem stability or the concept that succession was inherently driven to maximize some features (total biomass, biotic diversity, ratios of energy, or material transfers, etc.). These latter considerations can be seen as a logical continuation of Clements' concept that succession was progressive in its direction.

Odum's 1969 paper generated and still generates strong interest in the scientific community and continues to be influential these 40 years later. (At this writing in the end of year 2010, the Odum paper has been cited over 1,700 times with no diminution in citation rate.) As one might expect given its essential Clementsian theme updated with central topics in ecosystems ecology, the Odum's paper attracted several negative reviews from the ecologists of a Gleasonian persuasion, particularly given the ascendancy of ideas association with Gleason's papers on succession [6]. Nevertheless, Odum's work has had a strong influence on the direction of sustainability science. Its rules on how change should progress toward appropriate natural goals provide guidelines to how ecological systems might be managed toward naturally appropriate goals. Its dogmatism makes it a target for criticism but also gives it an influential position in those wrestling with the difficult problem in a dynamical, strongly human-altered planet.

Sustainability and Vegetation Dynamics

In the historical review above, one finds succession to be a complex concept with different interpretations as to its causes and attributes over time. Ecologists such as F.E. Clements and E.P. Odum viewed succession as a process that drove change in vegetation everywhere. The generality of their application invited the inspection of local examples and the discovery of "exceptions to the rule" that could whittle away at the claimed generality. They also were drawn to applying these principles in understanding human changes on landscapes. Other ecologists such as H.A. Gleason and probably the majority of modern ecologists active in this area of research today developed a much more

diffuse notion of succession. How do these ideas contribute to sustainability sciences and how can they be applied?

Clementsian Succession and Sustainability Science

The importance of succession for sustainability science can be illustrated by two examples: The first is from the historical role of succession theory in land-use policy in the United States; the second involves the incorporation of succession theory into environmental policy. Both examples favor Clementsian successional approaches in application in sustainability science. A following section will discuss the application of more Gleasonian approaches to sustainability science.

Succession and the United States Land-use Policy

The importance of succession theory in the land-use policy of the United States can be illustrated in human connections associated with F.E. Clements. Clements' doctoral advisor at the University of Nebraska was Charles Edwin Bessey, a leading botanist of his era. Bessey came to the University of Nebraska in 1884. He was a leading college administrator; he wrote high school and college textbooks; he served as editor for the journal *Science*. He produced a remarkable set of students: the Pulitzer Prize winning novelist, Willa Cather; Frederic Clements and his wife, Edith; and Roscoe Pound, the leading jurist of his time. Pound and Clements wrote several papers together. Pound went on to the faculty at Harvard University, became Dean of the Harvard Law School, and eventually served as an advisor to President Franklin Roosevelt involved in the structuring of the "New Deal" reformulation of the structure of American government. Among the "New Deal" legislation came administrative structures and regulations for land-use policies, conservation and land protection, soil conservation, and other topics which nowadays would clearly fall in the rubric of sustainability. Pound eventually wrote an obituary/recollection for F.E. Clements [44].

The connections among these remarkable individuals from American heartlands are indicators of a deeper relationship between the ecology of Clements, his associates and students, and land-use policies in the United States. The Great Depression of the 1930s in America was acerbated by a collapse of agriculture in the "Dust Bowl" states through the middle of the

nation. A complex combination of persistent land abuse, a climatic anomaly in the form of a drought, and bank collapse accompanying the collapse of family agriculture turned dryland farmers into unemployed refugees. The universities of the Midwest and the emergence of focused studies, such as the "Botanical Seminar" of Bessey at Nebraska (which had collected the Clements, Pound, and other students), had been in part a larger problem of understanding the difficulties in agriculture in the United States as farming moved westward from the forest of the East to the central grasslands [45]. Range management as an agricultural science grew from this tradition and particularly from the work of Clements and his colleagues [46, 47]. Thus, the emerging science of range management developing at the universities at the center of an ecological disaster involving the dynamics of grassland ecosystems produced information and applications on what could and should be done to restore stability. Political and policy advisors, such as Roscoe Pound, were in a position to translate this science and to put into laws concerning the sustainable use of grasslands, rangelands, and other ecosystems – in American and to a degree, worldwide.

Clarity of Clementsian Succession and the Formulation of Policy

Environmental policy like most political processes eschews complexity. One of the features of modern is the need for "elevator speeches" – summaries of complex issues that can be delivered to a listening decision-maker in the length of time one might have riding an elevator to a meeting. These and the even shorter snippets of information present to the press convey the idea that good ideas necessarily can be expressed succinctly. Certainly the Principle of Parsimony or "Ockhams Razor" in the sciences, the idea that one should select the hypothesis that makes the fewest additional assumptions when hypotheses are otherwise equal and a founding concept in modern science, lends itself in some interpreters to this point of view. Readers of F.E. Clements works are almost always impressed by the formidable richness of details in his writings, which he produced to support relatively clear general statements as to how succession as a unified process works. E.P. Odum, in some sense an inheritor and elaborator of Clements concepts, wrote using rich analogies to convey simple principles that operate in complex systems.

Clements' and Odum's writings share the ability to convey complexity with simple principles has served to make succession a tangible concept for policy makers and to illustrate the application of these principles. These "ecological generalizations as principles" have been widely used to impress upon the general public an appreciation of a unified ecological view of succession. The seemingly less-focused, special-case orientation of Gleason and subsequent ecologists seems complicated by exceptions. Range management research, which today still has a Clementsian theme and which has been successful in producing widely applied and useful policies for sustainable land management, is a tangible example that this approach can work. It is being challenged with its application in developing nations and in areas with old soils and variable climates but it remains the central paradigm [48].

Even though most modern ecologist would see Gleasonian views of succession as most appropriate, the findings do not always lend themselves to the "elevator-speech" test. One might ask, "What does an application of Gleasonian succession theory in sustainability look like in its application?" The answer is in modern forest and the very root of the origin of the word sustainability. This is the topic of the following section.

Nachhaltigkeit, Sustainability and Modern Forestry

Throughout Europe in Medieval times there was a substantial clearing of forests and extensive regional deforestation [49]. As forests, conflicts over the use products of forests intensified. This eventually produced a class conflict and subsequently laws against poaching of animals, thieving of wood, and proscriptions against public use of forests, in general. In the mid-eighteenth century, a forest management concept called "Nachhaltigkeit" (translated into English as the neologism, "sustainability") was developed by the Germans. From about 1,800, this new forestry practice spread over Europe, particularly Northern Europe. Nachhaltigkeit involved detailed determination of how to best manage forests to produce wood and other goods in perpetuity. In German, the root Nachhaltigkeit alone means nothing more than having a lasting or sustained effect.

Nachhaltigkeit involved detailed determination of how to best manage forests to produce wood and other goods. Essential to this objective was manipulating density by spacing trees on a given site, either by planting trees or by thinning a naturally regenerated stand of trees following a timber harvest or a natural disturbance. Additionally, one needed to determine how long one should wait before harvesting a stand of trees and then planting a new stand. This spacing/length-of-rotation problem had long been solved for crop plants through experimentation and observation. To produce trees as long-lived crops, elaborate long-term data collection started on the height; size in diameter; amount of wood; size of crowns in forest stands of different densities at sites with different environmental conditions. Eventually, a forest modeling concept called the "yield table" approach developed and became the signature of modern forestry. Nowadays, some of the historical forest data-sets used in yield tables have grown to 200+ years of continual record [50].

On a given kind of site (same soils, same rainfall, etc.), trees growing in even-aged, single-species stands (such as forest plantations) tend to grow to the same height at a given age, regardless of density of trees [51, 52]. At low densities, trees have with large diameters and crowns and on an equivalent high-density location, the trees would have small diameters and crowns – but the heights of the trees would be the same in both cases. In a yield table, decades and sometimes centuries of forest stand data are arranged by the height of the trees at a given location reach at a given age, usually the typical age of tree harvest. The tree height at this standard age is called a "site index" and is used to signify the overall quality of a location for growing trees. A site index is the canopy height one would expect a forest to attain in a reference length of time. If, for example, the dominant trees in the canopy of a forest attain a height of 30 m in 50 years (the reference time) then the forest would have a site index of 30. Sites with larger site indices would reach the 30 m height earlier than 50 years; sites with lower site indices would reach 30 m later (if at all). Using yield tables, one can assess the volume of wood in managed stands of equivalent heights on different sites.

Site index is clearly defined in terms of the basic data that goes into a yield table and it can be directly determined by measuring the heights of trees on even-aged

stands at the reference age. Associating site index with actual plots of land is a learned skill and an art at the same time. A capable site surveyor can judge site index by reconnaissance of land in a particular region and can make a good wage practicing this trade. Along with such arcane practices as ax-throwing, log-rolling, and tree-felling, forestry schools have regular intercollegiate competitions of judging site indices among their students as part of "Forestry Field Day" celebrations. At the edge of most universities with a college of forestry, one finds plots of trees planted at different densities for field teaching on the calibration of yield tables. The yield table concept is the quantitative basis of modern forestry.

One significant observation found in yield tables is that the rate of mortality of trees in stands undergoing natural thinning is highest in stands with the highest site indices. Faster growing trees suppress their competitors at a higher rate than slower growing trees on low site-index locations. The trade-off between rapid tree growth versus high mortality on good sites is the basis of the development of thinning strategies for managed stands and for schemes to plant forest plantations at particular spacing of trees. Despite their elevated mortality, sites with high site indices are more productive than low site-index locations.

Nachhaltigkeit was fueled by social unrest. It resulted in laws and regulation based on scientific research and experimentation with inspiration stemming from a shared political necessity to better manage the forests of Europe. Nachhaltigkeit was also the origin of the initial use in English of the word, sustainability. Its development into modern forestry practice was an ongoing process that continues today. Modern forestry is significant in sustainability science as a first real-world application. Modern forestry and its Nachhaltigkeit origins also represent a model of Gleasonian approaches to forest dynamics into sustainability science. It focuses on individual organisms and their relationships with one another and with their environment. One of the significant implications is that quantitative ecology and in particular, ecological models, likely could have a significant role in sustainability science. Individual-based vegetation dynamics models have a logical intellectual inheritance from forestry studies and are an example of tools for the application of Gleasonian successional concepts. These are the topics of the next section.

Individual-Based Models: Computer Descendants of Gleasonian Succession

Currently, several factors are simultaneously arising that challenge the application of successional theory in sustainability science. Climate conditions are changing, perhaps with a greater rate of change than seen in the past [53]. The carbon dioxide composition of the atmosphere has been altered by human activities with potential effects on plant processes particularly photosynthesis and the efficiencies of plant water-use [54]. This issue was addressed several decades ago by foresters who realized that the elaborate calibrations used to develop long-term data on forest change were subject to less accurate predictions if selected genetic strains of trees were used, if forests were fertilized or if there were significant variations in climate.

Impressed by the power of computers and interested in developing a methodology to use highly detailed computer models to simulate the spatial relations among thinning of trees, tree growth, and spatial arrangement of trees, foresters developed individual-based dynamic models of forests (mostly for commercial forests) in the mid-1960s. Their computational innovation was being independently paralleled in other fields, notably astronomy, physics, and several engineering sciences [55, 56]. The early individual-tree-based forest models were quite complex. For example, the competition among individual trees was typically simulated by crown interactions involving the 3-dimensional geometry of each individual tree crown for all the trees in a stand; [57] the growth of the tree trunks of each tree were often simulated at multiple heights; [58, 59] a tree's mortality was related to 2- or 3-dimensional crown pruning among trees [60].

For vegetation dynamics, the earliest individual-based model of forests [57] was developed for Douglas-fir (*Pseudotsuga menziesii*) applications and there were soon several similar models developed as doctoral dissertations at several Schools of Forestry [37]. These early individual-based models of forests were inordinately complex, perhaps because they were applied to tree-spacing problems in even-aged and single-species forests or forest plantations. However, a most detailed early model [61] with multiple tree species, tree interactions based on interacting three-dimensional crown geometry, mapping exact locations of trees, calculating

the dispersion of seed-fall based on seed morphology, etc., exceeds the complexity of almost all of the individual-based models of today.

As noted above, Watt noted that to understand succession one must account for the results of individual plants interacting with one another, but that it is “impractical at the individual level.” Twenty years later, the explosive expansion of computational power that still continues to date has made such computations more and more feasible. Early versions of these models in ecology were developed by population ecologists interested in including animal behavior in population models [62–64] and lead to a diverse array of applications for fish, insects, and birds [55]. An advantage of such models is that two implicit assumptions associated with traditional ecological modeling populations are not necessary, namely that:

1. The unique features of individuals (including their size and relative location) are sufficiently unimportant to the degree that individuals are assumed to be identical and
2. The population is “perfectly mixed” so that there are no local spatial interactions of any important magnitude [55].

As Watt noted, most ecologists are interested in variation between individuals (a basis for the theory of evolution and a frequently measured aspect of plants and animals) and appreciate spatial variation as being quite important. Assumption that this variation somehow is uniform seems particularly inappropriate for trees which are sessile and which vary greatly in size over their life span. This may be one of the reasons why tree-based forest models are among the earliest and most widely elaborated of this genre of models in ecology.

In 1972, Botkin and colleagues [65] produced an important simplification of the early forestry work (called the JABOWA model for the initials of its developers) and introduced this modeling approach to ecologists. They did not cite any of the antecedent forestry models and likely were not aware of them. The earlier forestry work remained sub rosa to most ecologists. The forestry models were mostly, but certainly not entirely, in sources not frequently read by forest ecologists. Additionally, some of the forestry models were seen as sufficiently useful that the models drifted

into the realm of “industrial secrets” as their developers found employment in the timber industry.

In 1980, a paper reviewing some of this work [66] coined the term, “gap model,” to describe this class of models. The gap-model designation was originally developed to emphasize that a principal simplifying assumption in these models (the assumption that the competition among individual trees on a small patch of land was homogeneous in the horizontal over a small area of land but spatially explicit in the vertical dimension) fitted well with the classic “gap dynamics” concept of A.S. Watt [39, 67]. At the time the term gap model was coined, Watt’s concept had been significantly reintroduced to American ecologists by Frank Bormann and Gene Likens [68, 69]. A generation of forest ecologists have made numerous extensions of “gap models” and the term, nowadays, refers to a broad class of individual-based models of forests and other ecosystems of a natural character (mixed-age, mixed-species, natural disturbance regimes, etc.).

An Application of an Individual-Based Model on Sustainability of Russian Forests under Climate Change

Global climate model simulations indicate that the Northern Hemisphere’s boreal forests and, in particular, the Siberian boreal forest zone, may not only respond to climate change but may affect the Earth’s climate through feedbacks involving changes in the regional surface albedo, the degree to which the surface reflects incoming radiation. Bonan and his colleagues [70] altered surface albedo in order to simulate the clearing of the boreal forest in the National Center for Atmospheric Research (NCAR)’s Community Climate System Model version 1.0 (CCSM1). This substantially cooled the Earth not only in the boreal zone but across the Northern Hemisphere. Betts [71] used the Hadley Center Atmosphere Model (HadAM3) to simulate the climatic consequences of albedo changes from growing more trees, worldwide. He found that the surface albedo changes associated with the growth of coniferous evergreen trees in boreal regions led to significant increases in the average global temperature. These increases were large enough to overshadow the effect of the carbon storage that occurred as a result of growing evergreen forest in that region.

Field observations provide further evidence that changes in the boreal forest may impact the global climate. Larch forest, dominated by both *Larix sibirica* and *L. gmelinii*, covers extensive regions in Siberia. Shifts from larch to dark-conifer forests, dominated by trees such as spruce or fir that are tolerant of higher temperatures, have been documented [72, 73]. Because larch is a deciduous conifer, this shift in forest composition would lead to the similar albedo changes as the evergreen tree growth simulation presented by Betts. This reduction of albedo associated with a documented shift in forest type (larch to dark conifer) indicates that warming temperatures may lead to a positive feedback response: a warmer climate accelerates the natural succession from larch to dark-conifer forest; the resultant albedo change promotes additional warming.

Dynamic vegetation models, specifically forest gap models, are ideally suited to an exploration of the impacts that climate change may have on the structure and composition of boreal forests and the existence of a climate/land cover feedback in this region. The FAREAST [74] model was run at a total of 2,083 sites across the former USSR. FAREAST uses monthly climate parameters derived from historical station data to compute daily temperature and update soil water [75]. In particular, at each site, the model's climate inputs are drawn from a statistical distribution of monthly values for minimum and maximum mean temperature and precipitation which is derived from 60 years of data recorded at local weather stations [76]. The model also uses values for soil field capacity and soil carbon and nitrogen [77].

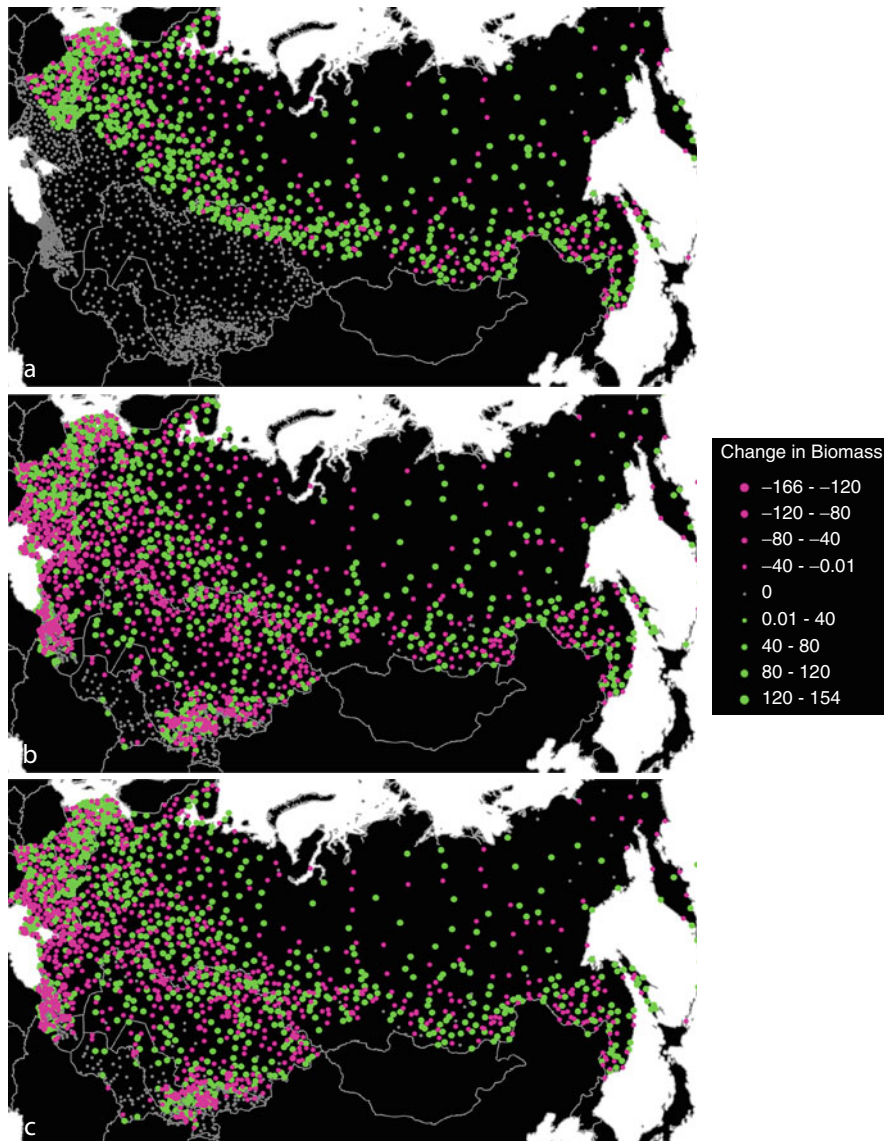
The birth, growth, and eventual death of individual trees are determined in response to local site parameters such as soil moisture and nutrient availability, which are updated annually with bio-environmental conditions, soil moisture, and available nutrients. Individual trees compete for light and nutrients with stochastic processes governing the birth and death of trees in a circular twelfth hectare plot, which approximates the size of a mature tree crown. Forty-four individual tree species are included in FAREAST simulations. These species represent the genera which dominate Northern Eurasian forests. Each species is characterized by 25 parameters which describe the species' fundamental silvics and determine which species

has an advantage in terms of competition for light or nutrients, or tolerance to lack of water. At each of the 2,083 sites, 200 independent twelfth hectare plots were simulated and then the modeled biomass values were averaged for each species in each year of the model run.

The overall response of Russian boreal forests to climate change when the effects of changes in temperature and precipitation are separated show higher average precipitation leads to increased biomass (Fig. 5a), lower average precipitation results in decreased biomass (Fig. 5b), and warming causes decreases in biomass for certain regions, though in parts of Siberia, where average temperatures are extremely low, warming induces an increase in forest biomass (Fig. 5c). There are also genera specific patterns in the shifts in biomass that occur across Russia. Specifically, there are different patterns of change for *Larix* spp. and *Pinus* spp. in response to temperature warming. Both genera display a decrease in biomass in western and southwestern Russia and the Russian Far East. The number of sites that experience a biomass decrease for *Larix* spp. is larger than the number of sites that show a decline in *Pinus* spp. In particular, the sites that show a decline in *Larix* spp. extend further northward in both European Russia and the Russian Far East. A more detailed analysis is required to determine whether these patterns are the result of a replacement of *Larix* spp. with *Pinus* spp.

Future Directions

Ecological succession has a complex and controversy rich history. It is significant over this history that the importance of successional knowledge as a prerequisite to the wise management of landscapes has not been a point of issue. The debates in succession are not about its importance. Rather, the debates were and are over how succession works, how it should be understood, and how it can be wisely applied. This is all the more so in the case of sustainability science as a user of successional theory and succession concepts. To sustainably manage the planet, it is an absolute prerequisite to have a capability to predict the dynamic change of the world's vegetation. To the extent that the understanding of succession and vegetation dynamics is not complete, so too will be the state of the understanding of sustainable management.



Ecological Succession and Community Dynamics. Figure 5

Decreasing biomass is shown in *pink* and increasing biomass is shown in *green*. Figure shows difference in magnitude of total forest biomass (tCha^{-1}) for year 200 between a baseline scenario with no change in climate, and: (a) climate scenario with a +10% precipitation showing an overall pattern toward increasing biomass; (b) climate scenario of -10% precipitation with pattern is toward decreasing biomass; (c) climate scenario with a 2°C increase with decreasing biomass in western and southwestern locations, but an inconsistent response across the region

Ecological models of vegetation dynamics have developed to a great degree over the past several decades. The capability to predict change in vegetation is much more today as it was 50 years ago. At the same time, the novel and large environmental changes that humankind is visiting upon the biosphere

(changes in the gasses of the atmosphere, in the diversity of regions, in the local to global climate, etc.) conspire to challenge models to predict outside the range of conditions in which they were developed. Extrapolation is always a difficult endeavor for life scientists.

One would expect sustainability scientists to become more dependent on models when conditions are novel. This novelty conspires to make models potentially less reliable. Sustainability has a word origin in the word, Nachhaltigkeit. That legacy grew into modern forestry with a quantitative capability to predict the future changes of forests under active management. In some senses, however, forestry with its emphasis on single-species, even-aged forest stands has developed the simple case and provided an example of how much effort it takes to develop such the simple case. The task of understanding the complex cases of natural ecosystems in the face of novel conditions is both daunting and essential. The real challenges to sustainability science is the development of meaningful policies and protocols when the essentials, such as succession theory, are themselves changing with new discovery.

Bibliography

Primary Literature

- Clements FE (1916) Plant succession: an analysis of the development of vegetation. Carnegie Institute, Washington, DC, Publication no 242
- Clements FE (1928) Plant succession and indicators. Wilson, New York
- Clements FE (1936) Nature and structure of the climax. *J Ecol* 24:252–284
- McCune B, Grace JB, Urban DL (2002) Analysis of ecological communities. MjM Software Design, Glenedenon, 304 pp
- Tansley AG (1935) The use and abuse of vegetational concepts and terms. *Ecology* 16:284–307
- Drury WH, Nisbet ICT (1973) Succession. *J Arnold Arbor* 54:331–368
- Gleason HA (1926) The individualistic concept of the plant association. *Bull Torrey Bot Club* 53:1–20
- Gleason HA (1939) The individualistic concept of the plant association. *Am Midl Nat* 21:92–110
- Rich PH (1988) The origin of ecosystems by means of subjective selection. In: Pomeroy LR, Alberts JJ (eds) Concepts of ecosystem ecology. Springer, New York, pp 19–28
- Egler FE (1954) Vegetation science concepts. I. Initial floristic composition – a factor in old-field vegetation development. *Vegetatio* 4:412–417
- McIntosh RP (1985) The background of ecology: concept and theory. Cambridge University Press, Cambridge, 383 pp
- Tansley AG (1947) Obituary notice. Frederic Edward Clements, 1874–1945. *J Ecol* 34:194–196
- White PS (1979) Pattern, process and natural disturbance in vegetation. *Bot Rev* 45:229–299
- Phillips J (1934) Succession, development, the climax and the complex organism: an analysis of concepts. I. *J Ecol* 22:554–571
- Phillips J (1935) Succession, development, the climax and the complex organism: an analysis of concepts. II. *J Ecol* 23:210–246
- Phillips J (1935) Succession, development, the climax and the complex organism: an analysis of concepts. III. *J Ecol* 23:488–508
- Heinselman ML (1981) Fire and succession in the conifer forests of northern North America. In: West DC, Shugart HH, West DC (eds) Forest succession: concepts and application. Springer, New York, pp 374–405
- Cattellino PJ, Noble IR, Slatyer RO, Kessell SR (1979) Predicting the multiple pathways of plant succession. *Environ Manage* 3:41–50
- Connell JH, Slatyer RO (1977) Mechanisms of succession in natural communities and their role in community stability and organization. *Am Nat* 111:1119–1144
- Glenn-Lewin DC, Peet RK, Veblin TT (eds) (1992) Plant succession: theory and prediction. Chapman and Hall, London
- Kessell SR, Potter MW (1980) A quantitative succession model for nine Montana forest communities. *Environ Manage* 4:227–240
- Weaver JE, Clements FE (1938) Plant ecology, 2nd edn. McGraw-Hill, New York, 520 pp
- McIntosh RP (1981) Succession and ecological theory. In: West DC, Shugart HH, Botkin DB (eds) Forest succession: concepts and application. Springer, New York, pp 10–23
- McIntosh RP (1975) H.A. Gleason – “Individualistic ecologist” 1882–1975; His contributions to theoretical ecology. *Bull Torrey Bot Club* 102:253–273
- Gleason HA (1910) The vegetation of the inland sand deposits of Illinois. *Bull Illinois State Lab Nat Hist* 9:21–174
- Gleason HA (1927) Further views on the succession concept. *Ecology* 8:299–326
- Gleason HA (1917) The structure and development of the plant succession. *Bull Torrey Bot Club* 44:463–481
- Cowles HC (1899) The ecological relations of the vegetation on the sand dunes of Lake Michigan. *Bot Gaz* 27:95–117, 176–202, 281–308, 361–369
- Cowles HC (1901) The physiographic ecology of Chicago and vicinity. *Bot Gaz* 31:73–108
- Adams CC (1935) The relation of general ecology to human ecology. *Ecology* 16:316–335
- Clements FE (1935) Experimental ecology in the public service. *Ecology* 16:342–363
- Transeau EN (1935) The prairie peninsula. *Ecology* 16:423–437
- Fuller GD (1935) Postglacial vegetation of the Lake Michigan region. *Ecology* 16:473–487
- Sears PB (1935) Types of North American pollen profiles. *Ecology* 16:488–499

35. Smuts JC (1926) Holism and evolution. Macmillan, London
36. Phillips J (1955) A tribute to Frederic E. Clements and his concepts in ecology. *Ecology* 35:114–115
37. Shugart HH (1998) Terrestrial ecosystems in changing environments. Cambridge University Press, Cambridge, 537 pp
38. Watt AS (1925) On the ecology of British beech woods with special reference to their regeneration. II. The development and structure of beech communities on the Sussex Downs. *J Ecol* 13:27–73
39. Watt AS (1947) Pattern and process in the plant community. *J Ecol* 35:1–22
40. Odum EP (1953) Fundamentals of ecology. W. B. Saunders, Philadelphia, 383 pp
41. Jordan CF (1986) Ecological effects of nuclear radiation. In: Oriens GH (ed) Ecological knowledge and environmental problem solving: concepts and case studies. National Academy Press, Washington, DC, pp 331–344
42. Zilversmit DB, Entenmann C, Fishler MC (1943) On the calculation of turnover time and turnover rate from experiments involving the use of labeling agents. *J Gen Physiol* 26:325–331
43. Odum EP (1969) The strategy of ecosystem development. *Science* 164:262–270
44. Pound R (1954) Frederic E. Clements as I knew him. *Ecology* 35:112–113
45. Arthur JC (1895) Development of vegetable physiology. *Science* 44:164–184
46. Tobey R (1976) Theoretical science and technology in American ecology. *Technol Cult* 17:461–534
47. Tobey R (1981) Saving the prairies: the life cycle of the founding school of American plant ecology 1895–1955. University of California Press, Berkeley
48. Washington-Allen RA, West NE, Douglas Ramsey R, Phillips DK, Shugart HH (2010) Retrospective assessment of soil stability on a landscape subject to commercial grazing. *Environ Monit Assess* 160:101–121
49. Williams E (2003) Deforesting the earth: from prehistory to global crisis. University of Chicago Press, Chicago, 689 pp
50. Pretzsch H (2009) Forest dynamics, yield and growth: from measurement to model. Springer, Berlin
51. Kimmins JP (1987) Forest ecology. Macmillan, Vancouver
52. Kira T (2001) Forest and environment: an approach to global environmental issues. Shin-Shisoshia, Tokyo
53. Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) (2007) Climate change 2007: the physical science basis. Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change. Cambridge University Press, Cambridge, UK/New York, 996 pp
54. Shugart HH, Woodward FI (2011) Global change and the terrestrial biosphere: achievements and challenges. Wiley-Blackwell, Oxford, 216 pp
55. Huston M, DeAngelis DL, Post WM (1988) New computer models unify ecological theory. *Bioscience* 38:682–691
56. DeAngelis DL, Gross LJ (eds) (1992) Individual-based models and approaches in ecology: populations, communities and ecosystems. Chapman and Hall, New York
57. Newnham RM (1964) The development of a stand model for Douglas-Fir. PhD thesis, University of British Columbia, Vancouver
58. Mitchell KJ (1969) Simulation of growth of even-aged stands of white spruce. *Yale Univ Sch Forest Bull* 75:1–48
59. Mitchell KJ (1975) Dynamics and simulated yield of Douglas-fir. *Forest Sci Monogr* 17:1–39
60. Hegyi F (1974) A simulation model for managing jack-pine stands. In: Fries J (ed) Growth models for tree and stand simulation. Royal College of Forestry, Stockholm, pp 74–90
61. Ek AR, Monserud RA (1974) FOREST: computer model for the growth and reproduction simulation for mixed species forest stands. Research report A2635, College of Agricultural and Life Sciences, University of Wisconsin, Madison
62. Holling CS (1961) Principles of insect predation. *Annu Rev Entomol* 6:163–182
63. Holling CS (1964) The analysis of complex population processes. *Can Entomol* 96:335–347
64. Rohlf FJ, Davenport D (1969) Simulation of simple models of animal behavior with a digital computer. *J Theor Biol* 23:400–424
65. Botkin DB, Janak JF, Wallis JR (1972) Some ecological consequences of a computer model of forest growth. *J Ecol* 60:849–872
66. Shugart HH, West DC (1980) Forest succession models. *Bioscience* 30:308–313
67. Watt AS (1925) On the ecology of British beech woods with special reference to their regeneration. II. The development and structure of beech communities on the Sussex Downs. *J Ecol* 13:27–73
68. Bormann FH, Likens GE (1979) Pattern and process in a forested ecosystem. Springer, New York
69. Bormann FH, Likens GE (1979) Catastrophic disturbance and the steady state in northern hardwood forests. *Am Sci* 67:660–669
70. Bonan GB, Pollard D, Thompson SL (1992) Effects of boreal forest vegetation on global climate. *Nature* 359:716–718
71. Betts RA (2000) Offset of the potential carbon sink from boreal forestation by decreases in surface albedo. *Nature* 408:187–190
72. Kharuk V, Ranson K, Dvinskaya M (2007) Evidence of evergreen conifer invasion into larch dominated forests during recent decades in Central Siberia. *Eurasian J Forest Res* 10: 163–171
73. Kharuk VI, Ranson KJ, Sergey TI, Dvinskaya ML (2009) Response of *Pinus sibirica* and *Larix sibirica* to climate change in southern Siberian alpine forest-tundra ecotone. *Scand J Forest Res* 24:130–139
74. Yan X, Shugart HH (2005) A forest gap model to simulate dynamics and patterns of Eastern Eurasian forests. *J Biogeogr* 32:1641–1658

75. Shuman JK, Shugart HH (2009) Evaluating sensitivity of Eurasian forest biomass to climate change using a dynamic vegetation model. *Environ Res Lett* 4:045024. doi:10.1088/1748-9326/4/4/045024
76. National Climate Data Center (NCDC) (2005) TD-9813 daily and sub-daily precipitation for the former USSR Version 1.0. NOAA National Climatic Data Center, Asheville
77. Stolbovoi V, McCallum I (eds) (2002) CD-ROM land resources of Russia. International Institute for Applied Systems Analysis and the Russian Academy of Science, Laxenburg
78. Van Cleve K, Viereck LA (1981) Forest succession in relation to nutrient cycling in the boreal forest of Alaska. In: West DC, Shugart HH, Botkin DB (eds) *Forest succession: concepts and application*. Springer, New York, pp 185–221

Ecological Systems, Introduction

RIK LEEMANS

Environmental Systems Analysis Group, Wageningen University, Wageningen, The Netherlands

It is estimated that there are roughly 8 million animal species, 600,000 fungi, 300,000 plants, and an unestimated number of microbial species on earth. It is further estimated that only about 25% of the animal, fungi, and plant species have been identified as yet. Further, the interactions among these species and their physical environment are known to an even lesser degree. Meanwhile, the earth's biota faces the prospect of climate change which may be either slow to manifest or extremely rapid, as we pass through a potential tipping point. In addition, human population is set to grow by two billion by 2045 from a present seven billion. This would certainly be rated as a population explosion, and these numbers together with expected (and hoped for) economic growth will stress the entire ecology of the earth. The question is how to cope with all of the above. That is actually the subject and goal of this entire *Encyclopedia* and our team of well over 700 scientists and engineers, and also the subject of this section from an ecological systems viewpoint. Our objective in this section is to provide a significant portion of the scientific and engineering basis of the systems ecology of the earth and to provide this in 14 detailed entries written at a level for use by university students through practicing professionals. Our

approach is to recognize that we cannot wait for a complete data set for the biota before assessing and planning and acting to preserve the ecological balance of the earth. Therefore, methodology is presented aimed at identifying the key interactions and environmental effects and enabling a systems level understanding even at our present state of factual knowledge.

For ease of use by students, each entry begins with a glossary of terms, while at an average length of 20 print pages each, sufficient detail is presented for utilization by professionals in government, universities, and industry. The reader is also directed to the closely related sections: ► [Earth System Monitoring, Introduction](#) and also ► [Transport and Fate of Chemicals in the Environment, Introduction](#).

Each of the entries is summarized below.

► [Ecosystems, Adaptive Management](#) – Adaptive management is an approach to natural resource management that emphasizes learning through management based upon the philosophy that knowledge is incomplete and much of what is thought to be known is actually wrong, but despite uncertainty managers and policymakers must act. This entry includes: Structure Decision Making, Participatory Active Adaptive Management, Adaptive Governance, and Adaptive Management and Law.

► [Marine Biogeochemistry](#) – The biogeochemistry of the world oceans has been studied for many decades, and major advances in understanding have been linked with development of new techniques and tools that allow the accurate representation of various organic and inorganic pools within the water. It has been shown that some critical bioactive compounds (containing the elements: carbon, nitrogen, phosphorus, oxygen) occur with particular atomic ratios to one another that are relatively invariant over space and time pools. The following cycles are presented in detail: carbon ocean Acidification, nitrogen and phosphorous, iron (including iron ocean fertilization), sulfur, and oxygen.

► [Atmospheric Biogeochemistry](#) – Biogeochemistry represents the interaction of biology, chemistry, and geology in the Earth system. For many processes, an understanding of biological uptake and emission, chemical processing, and geological sequestration is necessary to resolve the sources and sinks of a particular constituent. For example to discover the

sources and sinks of atmospheric carbon dioxide, it is important to understand how biota take up carbon dioxide and chemically convert the carbon to organic carbon, and then how this organic carbon is used either to produce energy by biota or is deposited to the land or ocean surface and can become sequestered in geological formations. Often when biogeochemistry is referred to, one refers to the nutrient cycling of important nutrients. Atmospheric biogeochemistry, as defined in its narrowest sense, is the flux of nutrients and pollutants important for biogeochemistry through the atmosphere. In its broadest sense, atmospheric biogeochemistry is the interaction of the atmosphere and atmospheric processes with the biogeochemistry of the earth system. This broader definition is used here.

► **Ecological Succession and Community Dynamics** – Ecological succession and the dynamic change of communities is one of the pillars of modern ecology and it has considerable influence on sustainability science. The processes that drive succession and the regularity of patterns of succession have been the topic of considerable debate over the past century – due in no small part to the importance of the concept to basic and applied ecology. Succession is at its simplest level the pattern of change in ecological communities and these changes can be read on the landscape if one knows regional history or conversely can be used to interpret a landscape's history.

► **Ecosystem Engineers, Keystone Species** – This contribution focuses on two ecological phenomena. The first is “keystone species” which is defined as a species (mostly of high trophic status) whose activities exert a disproportionate influence on the patterns of species occurrence, distribution, and density in a community. The second is the concept of “ecosystem engineers” defined as organisms that directly or indirectly modulate the availability of resources (other than themselves) to other species by causing physical state changes in biotic or abiotic materials. Models have shown that the loss of species with a large number of trophic interactions can trigger high numbers of secondary extinctions with serious consequences for species persistence. The importance of keystone species can also be strongly linked with ecosystem engineering. For example, the large impact of sea otters in kelp forest ecosystems results from the coupling between engineering effects and a trophic cascade.

► **Ecosystem Services** – This entry provides an overview of issues concerning the identification, biophysical and economic characterization, and safeguarding of ecosystem services. The concept of ecosystem services has a long written history, reaching back at least as far as Plato. This review history focuses on recent advances such as the Millennium Ecosystem Assessment, and advances in spatial modeling, economic valuation, and policy implementation. Examples of novel finance and policy mechanisms, including water funds, marine spatial planning, land-use planning and human development, and global policy efforts are presented. The entry concludes with a discussion of the largest research and implementation challenges in this field, identifying the issues that will frame the future growth of the concept of ecosystem services.

► **Ecosystems and Spatial patterns** – Ecological processes such as forest disturbances act on ecosystems at multiple spatial and temporal scales to generate complex spatial patterns. These patterns in turn influence ecosystem dynamics and have important consequences for ecosystem sustainability. Analysis of ecosystem spatial structure is a first step toward understanding these dynamics and the uncertain interactions among processes. Quantitative characterization of spatial patterns and their rates of change in natural environments are essential to understanding ecological processes and to inform sustainable management techniques that aim to minimize degradation and alteration of ecosystem dynamics. This entry covers: sources of heterogeneity, spatial analysis, spatial analyses within the ecosystem and spatial analyses among ecosystems.

► **Ecosystem Flow Analysis** – It is necessary to understand the ecosystem as a thermodynamic system and how the energy flows enter, interconnect, and disperse from the environmental system. Ecological network methodologies exist to investigate and analyze these flows. In particular, partitioning the flow into boundary input, non-cycled internal flow, and cycled internal flow shows the extent to which reuse and recycling arise in ecosystems. The intricate, complex network structures are responsible for these processes all within the given thermodynamic constraints. Design of sustainable human systems could be informed by these organizational patterns, in order to use effectively the energy available. This entry

demonstrates the need for flow analysis, provides a brief example using a well-studied ecosystem, and discusses some of the ecosystem development tendencies which can be addressed using ecosystem flow analysis.

► **Ecotones and Ecological Gradients** – Ecotones are areas of steep transition between ecological communities, ecosystems, or ecological regions (such as Mediterranean and desert). Ecotones often occur along ecological gradients. Such gradients are created as a result of spatial shifts in elevation, climate, soil, and many other environmental factors. Ecotones commonly coincide with areas of sharp climatic transition along environmental gradients. They occur at multiple spatial scales, from continental-scale transitions between major biomes to small-scale ecotones where local vegetation communities and microhabitats coincide. They show a diversity of boundary types that range from natural boundaries (e.g., altitudinal, latitudinal transitions) to human-generated ecotones (e.g., forest clear-cut edges or urban ecotones). Ecotones have been studied in the past four decades in an ecological context and in recent years are receiving increasing attention in the context of biodiversity conservation. Various studies have shown that species richness and abundances tend to peak in ecotonal areas, though exceptions to these patterns occur. Ecotones are “natural laboratories” for studying a range of evolutionary processes, such as the process by which new species form, also termed speciation. Recent evidence suggests that ecotones may serve as speciation centers. Some researchers argue that ecotones deserve high conservation investment, potentially serving as speciation and biodiversity centers. Because ecotones are often small in size and relatively rich in biodiversity, conservation efforts in these areas may prove to be an efficient and cost-effective conservation strategy.

► **Invasive Species** – This entry describes the causes and consequences of biological invasions, by synthesizing concepts from population biology, community ecology, evolution, biogeography, and conservation biology. First, the patterns and process of invasion are explored; then, some of its potential ecological and socioeconomic impacts are examined. Some major hypotheses and theoretical concepts explaining patterns of colonization and impact are presented. Next, management approaches to assessing,

preventing, and mitigating this problem are considered. The entry ends with a brief glimpse at some of the emerging issues that will likely be the foci of future research.

► **Landscape Ecology** – Landscapes are spatially heterogeneous areas characterized by diverse patches that differ in size, shape, contents, and history. When spatial heterogeneity is considered, scale matters and hierarchy emerges. Landscape ecology is the science and art of studying and improving the relationship between spatial pattern and ecological processes on a multitude of scales and organizational levels. In a broad sense, landscape ecology represents both a field of study and a scientific paradigm. As a highly interdisciplinary and transdisciplinary enterprise, landscape ecology integrates biophysical and analytical approaches with humanistic and holistic perspectives across natural and social sciences. Landscape ecology was initially developed in Europe. With theoretical developments in spatial ecology and technological advances in remote sensing and geospatial information processing, landscape ecology became an internationally recognized field of study in the 1980s. The most salient characteristics of landscape ecology are its emphasis on the pattern-process relationship and its focus on broad-scale ecological and environmental issues. Key research topics in landscape ecology include ecological flows in landscape mosaics, land use and land cover change, scaling, relating landscape pattern analysis with ecological processes, and landscape conservation and sustainability.

► **Species Competition and Predation** – Competition and predation are key interactions between species, and are major foci of thought and study in community ecology. They are believed to be major forces structuring natural communities, having critical roles in the determination of species diversity and species composition, and are regarded as important drivers of evolutionary processes. The relationships between the niches of different species determine how they interact through competition and predation, which then have key roles in assembly of local communities, and their reassembly following perturbations. Niches define patterns of linkages between species, their resources, and their natural enemies. These patterns include how linkages change over time, and between different spatial locations, and define

mechanisms by which similar species are able to coexist by their effects on competition and predation relationships. The human element in the environment has profound effects on these phenomena. Changing the environment shifts interactions between species, and profoundly modifies the structure of food webs. In the modern day, there is much community reassembly, potentially involving major shifts in competition and predation. Humans transport invasive species that act as predators, prey, and competitors with potentially major effects on the community reassembly process.

► **Species Diversity Within and Among Ecosystems** – Species diversity is a function of species richness, the number of species in a given locality and species evenness, the degree to which the relative abundances of species are similar. Commonly used quantification methods include constructing mathematical indices known as diversity indexes (the Shannon, Simpson, and Margalef indexes being the most widely used) or comparing observed patterns of species abundance to theoretical models. The scale of assessment may range from within a single site or habitat (known as α diversity) to the difference between two or more sites (β diversity); which can then be combined to give γ diversity – the diversity of the landscape. Early research tended to be concentrated on largely α diversity, but there has been a marked increase in studies investigating β and γ diversity during the last decade. The species richness aspect of diversity is often considered the iconic measure of biodiversity since it is well defined and aligns with our intuitive sense of the concept. As the species level is the level at which living organisms are most widely known, species have become the major taxonomic rank for describing biodiversity. Along with populations, it is also the level that most scientists, managers, and policy makers use when referring to biological diversity and it provides a useful means for both monitoring and priority setting. As the species level is one of three fundamental levels of biodiversity (the others are the genetic level and the ecosystem/community level), species diversity is thus an important component of conservation.

► **Urban Ecology** – Within the science of ecology, urban ecology is defined as the study of structure, dynamics, and processes in urban ecological systems. Urban ecology is the study of the relationships of human and nonhuman organisms in urban areas, the

interactions of these organisms with the native and built physical environment, and the effects of these relationships on the fluxes of energy, materials, and information within individual urban systems and between urban and nonurban systems. Urban ecology applies the methods and concepts of the biological science of ecology to urban areas, but requires and integrates with the concerns, concepts, and approaches of social sciences to produce a hybrid discipline. Urban ecological systems include individual organisms, populations, communities, and landscapes, as well as buildings and infrastructure. Urban ecology further recognizes specific urban ecosystems as a part of the global biogeochemical, economic, and human demographic system.

Ecosystem Engineers, Keystone Species

SANNE DE VISSER¹, ELISA THÉBAULT², PETER C. DE RUITER³

¹Community and Conservation Ecology group, University of Groningen, Groningen, The Netherlands

²CNRS, UMR 7618 “Biogéochimie et écologie des milieux continentaux”, Paris, France

³Biometris, Wageningen UR, Wageningen, The Netherlands

Article Outline

Glossary

Definition of the Subject

Introduction: Keystone Species and Ecosystem Engineers: Analysis of Concepts

Issue-1: How to Find Keystone Species and Ecosystem Engineers in Communities?

Issue-2: Usefulness for Management

Future Directions

Bibliography

Glossary

Connectance The proportion of possible ecological interaction links between species that are realized.

Ecosystem engineer An organism that creates or modifies its habitat [19].

Ecosystem functioning The way ecosystems work related to abiotic and biotic components, such as chemicals, water, soil, microbes, plants, and animals.

Keystone species A species that has a disproportionate effect on its environment relative to its biomass (Paine 1995).

Trophic level The position a species occupies in a food chain.

Definition of the Subject

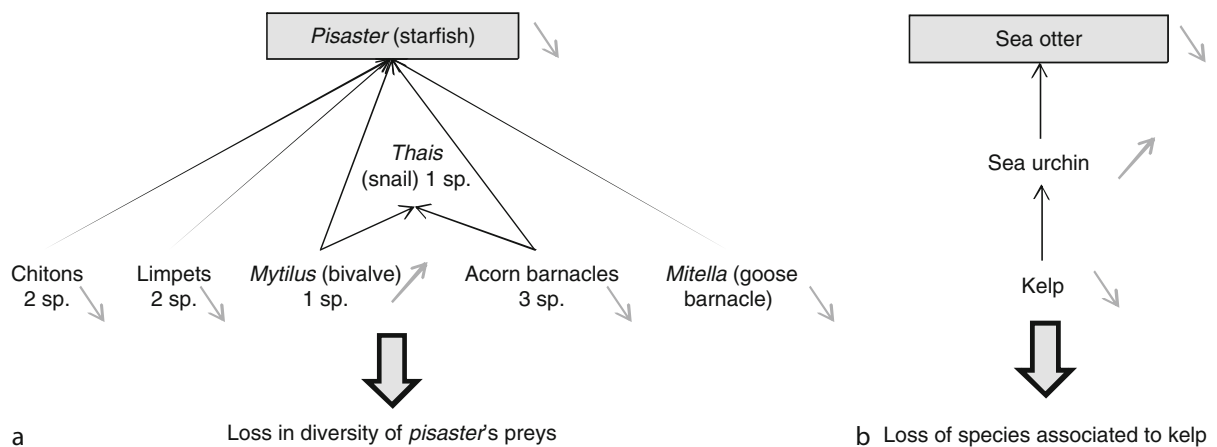
This entry focuses on two ecological phenomena. The first is “keystone species” which is defined by Paine [1] as a species (mostly of high trophic status) whose activities exert a disproportionate influence on the patterns of species occurrence, distribution, and density in a community. The second is the concept of “ecosystem engineers” defined by Jones et al. [2] as organisms that directly or indirectly modulate the availability of resources (other than themselves) to other species by causing physical state changes in biotic or abiotic materials.

Introduction: Keystone Species and Ecosystem Engineers: Analysis of Concepts

Paine’s definition of keystone species was inspired from the large effects of the removal of the carnivorous

starfish (*Pisaster ochraceus*) from intertidal habitat, which reduced prey species diversity due to intense competition from mussel prey [3], and represents now a classic textbook in ecology. The original keystone species concept of Paine [1, 4] thus identified a very specific mechanism: the top-down regulation of community structure and diversity by a top predator (Fig. 1). The concept of keystone species has been later extended to a broader definition and now includes any species whose effect on ecosystems is disproportionately large relative to its low biomass in the community as a whole [5]. Keystone species are thus species which have large effects on communities and ecosystems through many different processes such as trophic interactions, pollination, or habitat modification [6, 7]. Examples include rabbits that can increase abundance and diversity of lizards [8] and sea otters whose hunting in the late nineteenth century caused a population explosion of their sea urchin prey and consequent overgrazing of kelp which led to numerous extinctions of local species [6].

The concept of ecosystem engineering was proposed two decades later than the “keystone species” concept by Jones and colleagues [2]. They defined ecosystem engineers as “organisms that directly or indirectly modulate the availability of resources



Ecosystem Engineers, Keystone Species. Figure 1

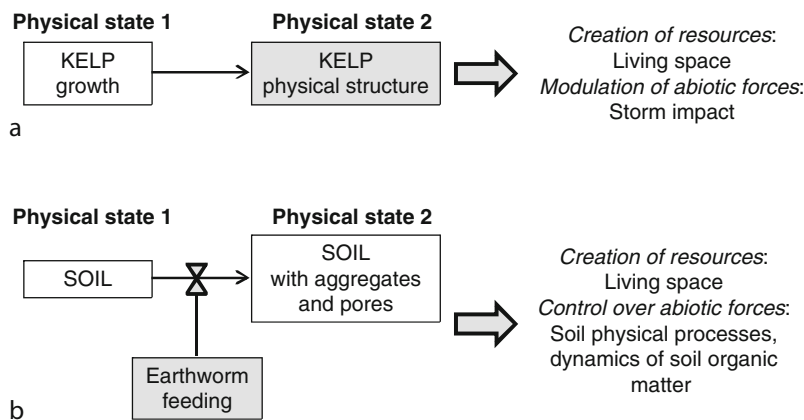
Two examples of keystone species impacts. **(a)** Effects of the removal of *Pisaster* on prey species diversity as a consequence of mussel population explosion. **(b)** Consequences of the removal of sea otters on species diversity due to overgrazing of kelp by sea urchins. Keystone species are represented in grey boxes. Small grey arrows indicate the direction of species abundance changes following the removal of the keystone species. The large grey arrows indicate the global consequences of keystone species loss on the ecosystem

(other than themselves) to other species by causing physical state changes in biotic or abiotic materials. In so doing they modify, maintain and/or create habitats” [2]. They further distinguish between two types of ecosystem engineers (Fig. 2): autogenic engineers that change the environment via their own physical structures, i.e., their living and dead tissues, and allogenic engineers that change the environment by transforming living or nonliving materials from one physical state to another via mechanical or other means. The idea that organisms can have important effects on abiotic processes occurring in the environment had been recognized before; indeed, Darwin devoted a whole book to the impact of earthworms on soil formation [9]. However, since the development of the concept of ecosystem engineer, engineering effects have been described for many organisms, from classic examples such as beavers, termites, or earthworms [10–12] to mollusks [13], fish [14], caterpillars [15], polychaete worms [16], grasses [17, 18], burrowing shrimp [19], ants [20], and many other species (see Table 1 of [21]).

Both the keystone species and the ecosystem engineer concepts point out to species which have

important effects in ecological communities and ecosystems. Although these concepts partly overlap – an ecosystem engineer can be a keystone species – they however insist on different aspects: the keystone concept focuses on species which have disproportionate effects on community structure and ecosystem functioning (“outcome focused” sensu [22]) whereas the ecosystem engineering concept considers organisms which influence the abiotic environment with consequences on other species and related ecosystem processes (“process focused” sensu [22]). These differences between the two concepts are reflected in the literature: these concepts generally appear in distinct studies as less than 5% of the studies on these topics refer to both ecosystem engineers and keystone species (source: ISI Web of Science).

The keystone species concept has been strongly related to food web theory since its first definition [1, 23]. In particular, the identification of keystone species in food webs is an important issue. Theoretical studies have tried to pin down the characteristics of keystone species through food web models [24, 25] and several indices based on food web topology have been developed to identify keystones [26]. Models have



Ecosystem Engineers, Keystone Species. Figure 2

Two examples of ecosystem engineering with kelp (a) and earthworms (b) as ecosystem engineers. Ecosystem engineering corresponds to changes in physical state (state 1–2) of biotic (i.e., kelp for a) or abiotic (i.e., soil for b) materials. Ecosystem engineers are represented in grey boxes. (a) Case of autogenic engineering, the engineer is part of the new physical state (via growth here). (b) Case of allogenic engineering, the new physical state is caused by the engineer (x, caused via feeding here), but the engineer is not part of the new physical state

shown that the loss of species with a large number of trophic interactions can trigger high numbers of secondary extinctions with serious consequences for species persistence; a result which highlights the potential keystone role of highly connected species in food webs [24, 27, 28].

In contrast, the concept of ecosystem engineering has been rarely related to food web studies. Recent studies acknowledge that ecosystem engineers may also play an important role in the network of trophic interactions but separating the trophic effects from the engineering effects to determine their relative importance is difficult [20, 29–31].

The importance of keystone species can also be strongly linked with ecosystem engineering. For example, the large impact of sea otters in kelp forest ecosystems results from the coupling between engineering effects and a trophic cascade [32]. In these ecosystems, kelp provides habitat for many species and dampens wave action; the keystone effect of sea otters is thus mediated through their indirect trophic effect on kelp densities which is a main ecosystem engineer.

Issue-1: How to Find Keystone Species and Ecosystem Engineers in Communities?

Keystone species and ecosystem engineers may affect ecosystem processes, such as nutrient cycling, and thereby ecosystem functioning. In the face of rapid biodiversity loss, a considerable amount of studies were dedicated to investigate a possible link between species richness and ecosystem function [33] and the threat of diversity loss on the loss of ecosystem services to man. First indications show positive relationships between species richness and ecosystem productivity, stability, and sustainability, with more species being able to fully and complementarily run ecosystem functions due to niche differentiation and facilitative interactions (reviewed by [34]). However, there is now a growing consensus that functional diversity, rather than species numbers per se, strongly determines ecosystem functioning [35]. This means that the presence of a particular species with specific traits may play a larger role in determining ecosystem function than merely the number of species [36]. The apparent diversity-ecosystem function relationship can thus be partly caused by a greater chance of an

influential species with particular traits being present in more diverse communities than in species-poor communities.

If it is possible to predict and identify a priori a set of species traits that determine keystone interactions in a system, this would greatly benefit management and conservation purposes. Species' traits determine how species contribute to ecosystem processes, so the presence and distribution of such traits can be utilized to indicate aspects of ecosystem functioning [37]. To identify keystone species various methods have been used ranging from experimental removal or addition manipulations to comparative studies and natural history observations [5]. Partly because of these methodological issues, identifying keystone species has so far proved elusive [5, 38] although some progress has been made and its concept now widely investigated in the context of complex ecological networks [25, 39–41]. Some examples of specific traits are for instance trophic level, body size, connectance, or traits concerning tolerance and resilience to disturbances. Organisms that influence their environment strongly and contribute disproportionately to the functioning of ecosystems often seem to occupy higher trophic levels in food webs [5]. Top predators have been described as highly interactive keystone species [42], have been shown to play an important role in stabilizing food webs [43], and play important roles in marine ecosystems [44] and terrestrial ecosystems [45].

Also, the loss of top predators has been linked to secondary extinctions [46, 47]. This has been attributed to their ecological role as suppressors of medium-sized predators (mesopredators) (e.g., [48, 49]) and generalist herbivores [50, 51]. In terrestrial ecosystems, organisms that influence their environment strongly also often seem to be large bodied (e.g., [52]). Larger bodied organisms require a high resource and energy use per individual [53, 54] and have greater mobility, home ranges, and longevity [55, 56] and, thereby, control more resources over greater and coarser spatial scales [52, 57]. It is also proposed that well-linked and interacting species as key interactors are more important for the community [28, 58–62]. This approach characterizes the interaction structure of species placed in an ecological network. Among plants, on the other hand, some studies have shown that species within the same functional types but with different requirements

and tolerances may provide insurance to the system in the form of long-term resilience against changes in environmental factors, such as global warming, grazing, drought or frost [35].

The latter example indicates that the keystone status of a species often appears to be context dependent, and may change with successional status, productivity, diversity, and other ecosystem traits [63]. It is therefore important to identify how the importance of traits that define keystone species change across a gradient of conditions, measuring environmental factors, community composition, trophic dynamics, and distribution of strong and weak links in the community (e.g., [24]). Without droughts, a specific plant species may not play an important role in maintaining community composition or ecosystem functioning. The Australian brushtail possum (*Trichosurus vulpecula*) may function as a keystone species in rata-kamahi forests by defoliating and killing canopy trees, but not in beech-dominated forests where floristic composition, but not forest structure, is typically affected [64–66]. A species in its native grounds may play no specific role in the system, whereas an invasive species may have devastating effects in the system it got introduced into, e.g., feral cats and rats on islands [67], Alewife (*Alosa pseudoharengus*) in nonnative freshwater lakes and ponds [68], and Cheatgrass (*Bromus tectorum*) in nonnative grasslands [69].

Identifying keystone species therefore is not without its problems. It is also important to notice that ecologically important species might not necessarily be the ones that are also considered important by traditional conservationists (i.e., rare species; [70]).

Issue-2: Usefulness for Management

Because of the limited resources available in comparison to conservation needs, it has been proposed to design protection of single species in the aim of indirectly protecting the regional biota. These “surrogate species” are roughly of three categories [71]: (1) flagships, charismatic species that attract public support; (2) umbrellas, species requiring such habitats that their protection might protect other species; and (3) biodiversity indicators, taxa whose presence may indicate high species richness. However the effectiveness of these policies has been questioned

and [70] suggested that single-species management might be more effective when directed toward keystone species. Indeed, the importance of keystone species and ecosystem engineers in communities make these species particularly important conservation targets, since the loss of these species can affect entire communities and ecosystems [72]. However, the main difficulty for applying these concepts to conservation issues lays on both the identification of keystone species and ecosystem engineers in communities and on the context dependence of their impacts, as discussed in the previous section. Thus, although these concepts appear relevant for conservation policies, it is still a long way from providing general and practical recommendations for conservationists and managers [71].

The concepts of keystone species and of ecosystem engineers could also be useful for other management issues in natural and anthropized ecosystems, such as for ecosystem restoration or agriculture. For example, in agro-ecosystems, several well-known ecosystem engineers have been used to improve soil fertility and crop yield. In some countries, farmers make use of the soil fertilizing effect of termites by spreading termite mound soil in their field [73]. Similarly, earthworm inoculation has generally positive effects on crop yield [74].

Future Directions

The notions of ecosystem engineers and keystone species have been playing prominent roles in ecology for several decades, still many questions and uncertainties ask for further investigations. Three of them are briefly described here. The first is how keystone roles and engineering effects are related to body-size: Do larger organisms have larger effects than smaller organisms? The second regards the context dependence of keystone roles and engineering effects: As the composition and structure of ecological communities are dynamic both in terms of species composition and species abundances, what does that imply for the role species have in communities and ecosystem functioning? The third line of research might be the most relevant for our society: How can the concepts for nature conservation, biodiversity protection, and the enhancement of environmental quality be used?

Bibliography

Primary Literature

- Paine RT (1969a) A note on trophic complexity and community stability. *Am Nat* 103:91–93
- Jones CG, Lawton JH, Shachak M (1994) Organisms as ecosystem engineers. *Oikos* 69:373–386
- Paine RT (1966) Food web complexity and species diversity. *Am Nat* 100:65–75
- Paine RT (1969b) The *Pisaster-Tegula* interaction: prey patches, predator food preference, and intertidal community structure. *Ecology* 50:950–961
- Power ME, Tilman D, Estes JA, Menge BA, Bond WJ, Mills LS, Daily G, Castilla JC, Lubchenco J, Paine RT (1996) Challenges in the quest for keystones. *Bioscience* 46:609–620
- Estes JA, Palmisano JF (1974) Sea otters: their role in structuring nearshore communities. *Science* 185:1058–1060
- Terborgh J (1986) Keystone plant resources in tropical forest. In: Soulé ME (ed) *Conservation biology*. Sinauer Associates, Sunderland, pp 330–344
- Bravo LG, Belliure J, Rebollo S (2009) European rabbits as ecosystem engineers: warrens increase lizard density and diversity. *Biodivers Conserv* 18:869–885
- Darwin CR (1881) The formation of vegetable mould through the action of worms, with observations of their habits. John Murray, London
- Dangerfield JM, McCarthy TS, Ellery WN (1998) The mound-building termite *Macrotermes michaelseni* as an ecosystem engineer. *J Trop Ecol* 14:507–520
- Eisenhauer N (2010) The action of an animal ecosystem engineer: identification of the main mechanisms of earthworm impacts on soil microarthropods. *Pedobiologia* 53: 343–352
- Wright JP, Jones CG, Flecker AS (2002) An ecosystem engineer, the beaver, increases species richness at the landscape scale. *Oecologia* 132:96–101
- Gutiérrez JL, Jones CG, Strayer DL, Iribarne OO (2003) Mollusks as ecosystem engineers: the role of shell production in aquatic habitats. *Oikos* 101:71–90
- Flecker AS, Taylor BW (2004) Tropical fishes as biological bulldozers: density effects on resource heterogeneity and species diversity. *Ecology* 85:2267–2278
- Lill JT, Marquis RI (2003) Ecosystem engineering by caterpillars increases insect herbivore diversity on white oak. *Ecology* 84:682–690
- Schwindt E, De Francesco C, Iribarne OO (2004) Individual and reef growth of the non-native reef-building polychaete *Ficopomatus enigmaticus* in a south-western Atlantic coastal lagoon. *J Mar Biol Assoc UK* 84:987–93
- Perelman SB, Burkart SE, León RJC (2003) The role of native tussock grass (*Paspalum quadrifarium* Lam.) in structuring plant communities in the Flooding Pampa grasslands, Argentina. *Biodivers Conserv* 12:225–238
- Fogel BN, Crain CM, Bertness MD (2004) Community level engineering effects of *Triglochin maritima* (seaside arrowgrass) in a salt marsh in northern New England, USA. *J Ecol* 92:589–597
- Berkenbusch K, Rowden AA (2003) Ecosystem engineering – moving away from ‘just-so’ stories. *New Zeal J Ecol* 27:67–73
- Sanders D, van Veen FJF (2011) Ecosystem engineering and predation: the multi-trophic impact of two ant species. *J Anim Ecol* 80:569–765
- Wright JP, Jones CG (2004) Predicting effects of ecosystem engineers on patch-scale species richness from primary productivity. *Ecology* 85:2071–2081
- Wright JP, Jones CG (2006) The concept of organisms as ecosystem engineers ten years on: progress, limitations, and challenges. *BioScience* 56:203–209
- De Ruiter PC, Wolters V, Moore JC, Winemiller KO (2005) Food web ecology, playing Jenga and beyond. *Science* 309:68–71
- Christianou M, Ebenman B (2005) Keystone species and vulnerable species in ecological communities: strong or weak interactors? *J Theor Biol* 235:95–103
- Libralato S, Christensen V, Pauly D (2006) A method for identifying keystone species in food web models. *Ecol model* 195:153–171
- Jordán F, Liu W-C, Davis AJ (2006) Topological keystone species: measures of positional importance in food webs. *Oikos* 112:535–546
- Solé RV, Montoya JM (2001) Complexity and fragility in ecological networks. *P Roy Soc B-Biol Sci* 268:2039–2045
- Dunne JA, Williams RJ, Martinez ND (2002) Network structure and biodiversity loss in foodwebs: robustness increases with connectance. *Ecol Lett* 5:558–567
- Wilby A, Shachak M, Boeken B (2001) Integration of ecosystem engineering and trophic effects of herbivores. *Oikos* 92: 436–444
- Zhang YX, Richardson JS, Negishi JN (2004) Detritus processing, ecosystem engineering, and benthic diversity: a test of predator-omnivore interference. *J Anim Ecol* 73:756–766
- De Visser SN, Freymann BP, Olff H (2011) The Serengeti food web: empirical quantification and analysis of topological changes under increasing human impact. *J Anim Ecol* 80:484–494
- Jones CG, Lawton JH, Shachak M (1997) Positive and negative effects of organisms as physical ecosystem engineers. *Ecology* 78:1946–1957
- Schwartz MW, Brigham CA, Hoeksema JD, Lyons KG, Mills MH, van Mantgem PJ (2000) Linking biodiversity to ecosystem function: implications for conservation ecology. *Oecologia* 122:297–305
- Giller PS, O'Donovan G (2002) Biodiversity and ecosystem function: do species matter? *Biol Environ* 102B:129–139
- Diaz S, Cabido M (2001) Vive la difference: plant functional diversity matters to ecosystem processes. *Trends Ecol Evol* 16:646–655

36. Sala OE, Lauenroth WK, McNaughton SJ, Rusch G, Zhang X (1996) Biodiversity and ecosystem functioning in grasslands. In: Mooney HA et al (eds) *Functional role of biodiversity: a global perspective*. Wiley, Chichester, pp 129–149
37. Bremner J (2008) Species' traits and ecological functioning in marine conservation and management. *J Exp Marine Biol Ecol* 366:37–47
38. Menge BA, Berlow EL, Balchette CA, Navarrete SA, Yamada SB (1994) The keystone species concept: variation in interaction strength in a rocky intertidal habitat. *Ecol Monog* 64: 249–286
39. Brose U, Berlow EL, Martinez LD (2005) Scaling up keystone effects from simple to complex ecological networks. *Ecol Lett* 8:1317–1325
40. Mills LS, Soulé ME, Doak DF (1993) The keystone-species concept in ecology and conservation. *Bioscience* 43:219–224
41. Okey TA, Banks S, Born AF, Bustamante RH, Calvopiña M, Edgar GJ, Espinoza E, Fariña JM, Garske LE, Reck GK, Salazar S, Shepherd S, Toral-Granda V, Wallem P (2004) A trophic model of a Galapagos subtidal rocky reef for evaluating fisheries and conservation strategies. *Ecol Model* 172:383–401
42. Soulé ME, Estes AJ, Miller B, Honnold DL (2005) Strongly interacting species: conservation policy, management and ethics. *BioScience* 55:168–176
43. Rooney N, McCann K, Gellner G, Moore JC (2006) Structural asymmetry and the stability of diverse food webs. *Nature* 442:265–269
44. Boyd I, Wanless S, Camphuysen CJ (eds) (2006) *Top predators in marine ecosystems*. Cambridge University Press, Cambridge, UK
45. Johnson CN (2010) Red in tooth and claw: how top predators shape terrestrial ecosystems. *J Anim Ecol* 79:723–725
46. Borrvall C, Ebenman B (2006) Early onset of secondary extinction in ecological communities following the loss of top predators. *Ecol Lett* 9:435–442
47. Henke SE, Bryant FC (1999) Effects of coyote removal on the faunal community in western Texas. *J Wildlife Manage* 63:1066–1081
48. Crooks JA, Soulé ME (1999) Lag times in population explosions of invasive species: causes and implications. In: Viken A (ed) *Invasive species and biodiversity management*. Kluwer, Dordrecht, pp 103–126
49. Letnic M, Koch F, Gordon C, Crowther MS, Dickman CR (2009) Keystone effects of an alien top-predator stem extinctions of native mammals. *Proc Roy Soc Biol Sci* 276:3249–3256
50. McLaren BE, Peterson RO (1994) Wolves, moose, and tree rings on isle royale. *Science* 266:1555–1558
51. Ripple WJ, Beschta RL (2003) Wolf reintroduction, predation risk, and cottonwood recovery in Yellowstone National Park. *Forest Ecol Manag* 184:299–313
52. Owen-Smith RN (1988) *Megaherbivores: the influence of very large body size on ecology*. Cambridge University Press, Cambridge, UK
53. Bush GL (1993) A reaffirmation of Santa Rosalia, or why are there so many kinds of small animals? In: Lees DR, Edwards D (eds) *Evolutionary patterns and processes*. Academic, London, pp 229–249
54. Olff H, Ritchie ME, Prins HHT (2002) Global environmental controls of diversity in large herbivores. *Nature* 415:901–904
55. Haskell JP, Ritchie ME, Olff H (2002) Fractal geometry predicts varying body size scaling relationships for mammal and bird home ranges. *Nature* 418:527–530
56. Peters RH (1983) *The ecological implications of body size*. Cambridge University Press, Cambridge, UK
57. Ritchie ME, Olff H (1999) Spatial scaling laws yield a synthetic theory of biodiversity. *Nature* 400:557–560
58. Allesina S, Bodini A, Bondavalli C (2006) Secondary extinctions in ecological networks: bottlenecks unveiled. *Ecol Model* 194:150–161
59. Dunne JA, Williams RJ, Martinez ND (2004) Network structure and robustness of marine food webs. *Marine Ecol Prog Ser* 273:291–302
60. Estrada E (2007) Characterisation of topological keystone species: local, global and “meso-scale” centralities in food webs. *Ecol Complex* 4:48–57
61. Jordán F, Benedek Z, Podani J (2007) Quantifying positional importance in food webs: a comparison of centrality indices. *Ecol Model* 205:270–275
62. Jordán F, Liu W-C, Mike A (2009) Trophic field overlap: a new approach to quantify keystone species. *Ecol Model* 220:2899–2907
63. De Ruiter PC, Neutel A, Moore JC (1995) Energetics, patterns of interaction strengths, and stability in real ecosystems. *Science* 269:1257–1260
64. Batcheler CL (1983) The possum and rata-kamahi dieback in New Zealand: a review. *Pac Sci* 37:415–426
65. Rose AB, Pekelharing CJ, Platt KH (1992) Magnitude of canopy dieback and implications for conservation of southern rata-kamahi (*Metrosideros umbellata*-*Weinmannia racemosa*) forests, central Westland, New Zealand. *New Zeal J Ecol* 16:23–32
66. Wardle JA (1984) *The New Zealand beeches: ecology, utilization and management*. New Zealand Forest Service, Christchurch, p 447
67. Holdaway RN (1999) A spatio-temporal model for the invasion of the New Zealand archipelago by the Pacific rat *Rattus exulans*. *J Roy Soc New Zeal* 29:91–105
68. Madenjian CP, Pothoven SA, Dettmers JM, Holuszko JD (2006) Changes in seasonal energy density of alewife (*Alosa pseudoharengus*) in Lake Michigan after invasion of dreissenid mussels. *Can J Fish Aquat Sci* 63:1–12
69. Knapp PA (1996) Cheatgrass (*Bromus tectorum* L.) dominance in the Great Basin desert. *Global Environ Chang* 6:37–52
70. Simberloff D (1998) Flagships, umbrellas, and keystones: is single-species management passé in the landscape era? *Biol Conserv* 83:247–257

71. Boogert NJ, Paterson DM, Laland KN (2006) The implications of niche construction and ecosystem engineering for conservation biology. *BioScience* 56:1–9
72. Crain CM, Bertness MD (2006) Ecosystem engineering across environmental gradients: implications for conservation and management. *BioScience* 56:211–218
73. Black HJ, Okwakol MJN (1997) Agricultural intensification, soil biodiversity and agroecosystem function in the tropics: the role of termites. *Appl Soil Ecol* 6:37–54
74. Fragoso C, Brown GG, Patron JC, Blanchart E, Lavelle P, Pashanasi B, Senapati B, Kumar T (1997) Agricultural intensification, soil biodiversity and agroecosystem function in the tropics: the role of earthworms. *Appl Soil Ecol* 6:17–35
75. Paine RT (1995) A conversation on refining the concept of keystone species. *Conserv Biol* 9(4):962–964

Ecosystem Flow Analysis

BRIAN D. FATH

Biology Department, Towson University, Towson, MD, USA

Advanced Systems Analysis Program, International Institute for Applied Systems Analysis, Laxenburg, Austria

Article Outline

Glossary
 Definition
 Introduction
 Investigation of Ecosystem Flows Using Network Analysis
 Example: Cone Spring Ecosystem Model
 Ecosystem Goal Functions
 Conclusions and Future Directions
 Bibliography

Glossary

Consumer Heterotrophic organism that consumes other organisms for their energy requirements.

Cycling The process by which energy or matter returns from its compartment of origin before exiting the system boundary.

Ecological goal function Tendency observed in the orientation or directional development of ecological systems.

Flow The transfer of energy or matter from one compartment in the system to another by active (feeding) or passive (death, egestion) means.

Network analysis A mathematical tool to study objects as part of a connected system and to identify and quantify the direct and indirect effects in that system.

Primary producer Photosynthesizing organism that captures external energy sources and brings it into the system as the basis for all subsequent thermodynamic activity.

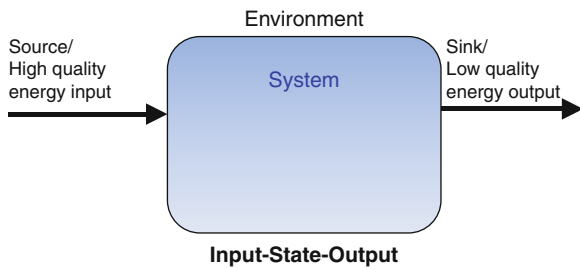
Thermodynamic system A bounded system defined by the quantities of energy and matter flowing through it.

Definition

Ecosystem services provide the basis for all human activity. Maintaining their sustained function is of critical concern to the issues of sustainability addressed here in this encyclopedia. At root, the ecosystem is a thermodynamic system receiving, collecting, transforming, and dissipating solar energy. The energy pathways are varied and complex and lead to the diversity of form and services available on the earth. Therefore, it is necessary to understand the ecosystem as a thermodynamic system and how the energy flows enter, interconnect, and disperse from the environmental system. Ecological network methodologies exist to investigate and analyze these flows. In particular, partitioning the flow into boundary input, noncycled internal flow, and cycled internal flow shows the extent to which reuse and recycling arise in ecosystems. The intricate, complex network structures are responsible for these processes all within the given thermodynamic constraints. Design of sustainable human systems could be informed by these organizational patterns, in order to use effectively the energy available. This article demonstrates the need for flow analysis, provides a brief example using a well-studied ecosystem, and discusses some of the ecosystem development tendencies which can be addressed using ecosystem flow analysis.

Introduction

Ecosystems, like all environmental systems, are open, thermodynamic systems (Fig. 1). They take in energy



Ecosystem Flow Analysis. Figure 1

Environmental systems are open systems, connecting to the environment through inputs of high-quality energy and discarding low-quality outputs

from an external energy source – almost entirely from solar energy, although geothermal or geochemical energy drives some systems. Ecosystem structure is built with the energy, and then the degraded energy is passed back to the environment. For some time period the “stored solar energy” persists in the forms perceived on the earth’s surface as biomass stores of living and nonliving organic material – such as all of us. In Frank Herbert’s novel *Dune*, he envisioned that on an arid planet an important functional role of each individual was “carrying” water in oneself. In a world dominated by thermodynamic constraints, such as ours, everyone is an energy carrier. These stores are temporary and fluid. In this perspective, one could substitute the current object-mode paradigm with a flow-based, rheomode. In other words, *all stocks are flows*. The paper that one writes on now is only an ephemeral stage of the energy which started in the solar reactions, traveled to earth, was captured by a photosynthesizing organism, converted into storage in the xylem of that organism, and harvested and transformed into the useable product that one now holds. But that is not the end. Over time, the paper will slowly degrade or decompose, or perhaps the energy release will be sudden through combustion. In any case, the objects held are transitory states in a long-term dynamic from (energy) source to (energy) sink. Changing this view would help in appreciating better the difference between capital and income, because, for example, harvesting natural capital stock (i.e., a forest) into a flow (deforestation) is not equivalent yet treated as substitutable in the current accounts. The rheomode approach could help focus clearly on the difference and thus sustainability of stocks and flows.

Concerned about sustainability over a human time horizon, one must be aware of the constraints imposed by these sources and sinks. From the input side, clearly, it is necessary that one does not extract resources at a rate faster than they can regenerate. And concerning the output, the waste emissions should not exceed to the assimilative capacity of the local environment [1]. These are the most basic constraints imposed by open-system, thermodynamics. Humans have transformed the earth’s surface to maximize the capture of photosynthetic energy – think of the millennia over which the Chinese, Romans, Babylonians, etc., have manipulated and manicured the landscape for agricultural production. Still, these societies rose and fell within the solar energy domain. These societies collapsed if they overconsumed the base resources or if they polluted their local environments [2]. In addition to these persistent input–output constraints, there is a third sustainability consideration currently observed in the anthropocene, in that it is not only the input–output relations, but also the structure which is created. Modern infrastructure demands the continual input of high-quality (low entropy) energy of a form not naturally delivered by ecosystem services. Furthermore, the created structure locks us into the necessity of immense energy flows for maintenance. Perhaps an apt analogy can be offered through the Greek myth of Erysichthon, who was King of Thessaly. He angered the gods by cutting down a sacred tree, and as punishment was insatiably hungry. Importantly, the more he ate, the hungrier he became. Our infrastructure, like Erysichthon, does not sit idle but continually demands upkeep such that the more structure, the more resources are needed to support this structure. It is not just about the present flows, but also the life cycle debt commitment as a result of the structure. Today that energy debt is paid almost entirely in the usage of fossil fuels – a nonrenewable resource. The scale of human activity seen today is because the application of fossil fuels to substitute solar fuels has released humans from one of the long-standing constraints on growth. And as a result, humans have exploded across the landscape. This growth was easy when there was sufficient energy to add to the system. In fact, the first growth form is *boundary growth*, taking energy into the system, and storing it as biomass. As long as there is more energy available, the system growth can occur unbound. The second stage

of growth, *network and information growth*, is squeezing more utility out of the available source by coupling processes and improving efficiencies [3]. This can occur in parallel with boundary growth, but becomes a necessity when boundary growth is limited. In human systems, those immediate constraints are looming at least given the current fuel-mix options.

Natural ecosystems, dependent on the solar energy flows developed extremely complex and beautiful structures within these thermodynamic constraints. And it is a useful guide to learn from these systems as a more eco-friendly design is incorporated. Below, some of the ecosystem flow analysis methodologies are explored and applied.

Investigation of Ecosystem Flows Using Network Analysis

“There are no trash cans in nature.” This is a useful phrase reminding that waste from one entity is food/input for another. Energy, of course, has a higher dissipative factor in the reuse than material cycles such as nitrogen, phosphorus, and calcium, but still there is a complex network of pathways designed to utilize the energy available in natural ecosystems. In 1973, Hannon [4] introduced Leontief’s input–output methodology into ecology, applying it to the energy flow structure of an ecosystem. The ecosystem is represented by n compartments, and the energy flowing into compartments, within compartments and exiting compartments.

A network flow model is essentially an ecological food web (energy–matter flow of who eats whom), which also includes energy input, and nonfeeding pathways such as dissipative export out of the system and pathways to detritus. The first step is to identify the system of interest and place a boundary (real or conceptual) around it. Energy–matter transfers within the system boundary comprise the network; transfers crossing the boundary are either input or output to the network, and all transactions starting and ending outside the boundary without crossing it are external to the system and are not considered. The energy inflows and intra-system flows can be considered the *production energy flow* and the flows with no consumers such as metabolic energy and exported biomass are the *respiration energy flow* [5].

The data required for ecological network analysis are as follows: for each compartment in the network, the biomass and physiological parameters, such as consumption (C), production (P), respiration (R), and egestion (E), must be quantified [6]. Furthermore, the diet of each compartment must be apportioned amongst the inputs from other compartments (consumption) in the network. This apportionment of “who eats whom and by how much” can be depicted in a dietary *flow matrix*, F , where energy flows from column elements j to row elements i . For all compartments, inputs should balance outputs ($C = P + R + E$) in accordance with the conservation of matter and the laws of thermodynamics.

The sum of the flow matrix elements, f_{ij} , gives the total inflow to compartmental i such that:

$$T_{i,in} = \sum_{j=1}^n f_{ij} + z_i$$

where z_i is the boundary flow into i . The outflow from i can be expressed as:

$$T_{i,out} = \sum_{j=1}^n f_{ji} + y_i$$

where y_i is the boundary outflow from i . At steady state, a necessary condition for the network flow analysis, $T_{i,in} = T_{i,out}$ and one compartmental throughflow vector can be written as $T = (T_i)$. The total system throughflow (TST) is given by the sum of the compartmental throughflows:

$$TST = \sum_{j=1}^n T_j$$

The motivation for flow partitioning begins with nondimensional flow intensities (i.e., throughflow-specific flows) which result when flows are divided by throughflows of originating compartments: $g_{ij} = f_{ij}/T_j$. The elements of matrix $G = (g_{ij})$ give the dimensionless transfer efficiencies corresponding to each direct flow, f_{ij} . Powers G^m of this matrix give the indirect flow intensities associated with paths of lengths $m = 2, 3, \dots$. Due to dissipation, flow along these indirect paths approaches zero as $m \rightarrow \infty$ so that the power series $\sum_{m=0}^{\infty} G^m$ representing the sum

of the initial, direct, and indirect flows converges to an integral flow intensity matrix, $\mathbf{N}$:

$$\underbrace{\mathbf{N}}_{\text{integral}} = \underbrace{\mathbf{I}}_{\text{initial}} + \underbrace{\mathbf{G}}_{\text{direct}} + \underbrace{\mathbf{G}^2 + \mathbf{G}^3 + \dots + \mathbf{G}^m + \dots}_{\text{indirect}} \\ = (\mathbf{I} - \mathbf{G})^{-1}$$

$\mathbf{N}$ maps the steady-state input vector $\mathbf{z}$ into the steady-state system throughflow vector:

$$\mathbf{T} = \mathbf{N}\mathbf{z} = (\mathbf{I} + \mathbf{G} + \mathbf{G}^2 + \mathbf{G}^3 + \dots + \mathbf{G}^m + \dots)\mathbf{z}$$

Term by term, flow intensities $\mathbf{G}^m$ of different orders m are propagated over paths of different lengths m . The first term, $\mathbf{I}$, brings the input vector $\mathbf{z}$ across the system boundary as input z_j to each initiating compartment, j . The second term, $\mathbf{G}$, produces the first-order direct transfers from each j to each i in the system. The remaining terms where $m > 1$ define m^{th} order indirect flows associated with length m paths. As stated before, these go to zero in the limit as $m \rightarrow \infty$, which is necessary for series convergence. This demonstrates that each “direct” flow f_{ij} at steady state is actually composed of flow elements of all orders, $m = 1, 2, \dots$. In fact, a major result of this flow analysis is that indirect flows can dominate direct flows: $\sum_{m=2}^{\infty} \mathbf{G}^m > \mathbf{G}$. In the above developments $\mathbf{F}$, $\mathbf{T}$, $\mathbf{z}$, and $\mathbf{y}$ represent matter or energy fluxes, and $\mathbf{G}$ and $\mathbf{N}$ are dimensionless intensive flows.

Finn [7] developed a cycling index using this basic approach and, Higashi et al. [8] described a three-mode partition of the flows, expanded by Fath et al. [9] into five modes (Table 1). Mode 0 is the boundary input into the system. Mode 1 accounts for all flow in which substance moves from node j to a terminal node i for the first time only without cycling. Mode 2 is flow cycled at terminal nodes i of each (i, j) pair. Mode 3 is component-wise dissipative flow in the sense that it exits from node i never to return again to i . Mode 4 is the boundary output from i constituting systemically dissipative flows exiting the system (Fig. 2). δ_{ij} is the Kronecker delta defined by $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$.

Note, the symmetry in that quantitatively Mode 0 = Mode 4, and Mode 1 = Mode 3. This is due to the conservation of mass/energy and at steady state what comes in must go out. Mode 2 represents the cycled flow which has additional impact on the system by staying in the system longer, increasing the residence

Ecosystem Flow Analysis. Table 1 Network representation of flow partitioning into five modes for any (i, j) pair in a system

	Pair-wise interaction	System-wide contribution
Mode 0 (boundary input)	$f_{j0}^{(0)} = z_j$	$f^{(0)} = \sum f_{j0}^{(0)}$
Mode 1 (first passage)	$f_{ij}^{(1)} = \left(\frac{n_{ij}}{n_{ii}} - \delta_{ij}\right)z_j$	$f^{(1)} = \sum \sum f_{ij}^{(1)}$
Mode 2 (cyclic)	$f_{ij}^{(2)} = \frac{n_{ij}}{n_{ii}}(n_{ii} - 1)z_j$	$f^{(2)} = \sum \sum f_{ij}^{(2)}$
Mode 3 (dissipation)	$f_{ij}^{(3)} = \left(\frac{n_{ij}}{n_{ii}} - \delta_{ij}\right)z_j$	$f^{(3)} = \sum \sum f_{ij}^{(3)}$
Mode 4 (boundary output)	$f_{0j}^{(4)} = y_j$	$f^{(4)} = \sum f_{0j}^{(4)}$

time, and returning to its source of emanation. Therefore, total system throughflow can be written as:

$$TST = f^{(0)} + f^{(1)} + f^{(2)} = f^{(2)} + f^{(3)} + f^{(4)}$$

And, on a nodal basis, throughflow is:

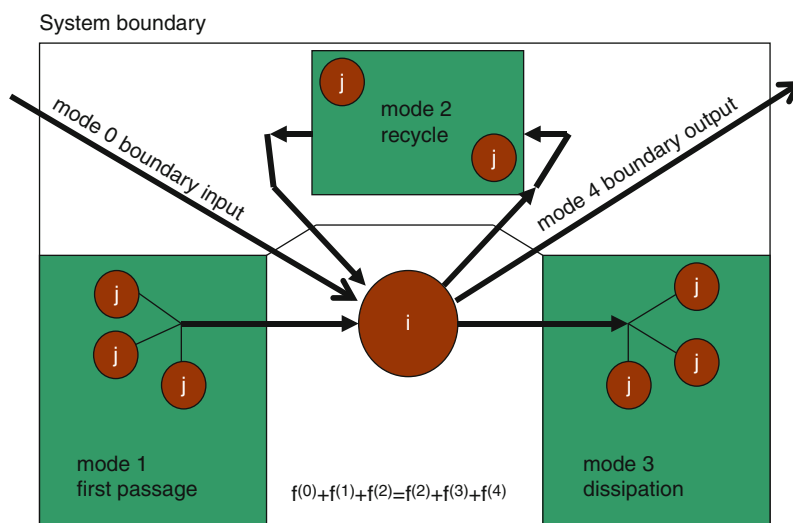
$$T_{ij} = f_{ij}^{(0)} + f_{ij}^{(1)} + f_{ij}^{(2)} = z_j \delta_{ij} + \left(\frac{n_{ij}}{n_{ii}} - \delta_{ij}\right)z_j + \frac{n_{ij}}{n_{ii}}(n_{ii} - 1)z_j \\ = \left(\delta_{ij} + \left(\frac{n_{ij}}{n_{ii}} - \delta_{ij}\right) + n_{ij} - \frac{n_{ij}}{n_{ii}}\right)z_j = n_{ij}z_j$$

The mode partition designation clearly shows the contribution of flow within the entire system of interactions.

Example: Cone Spring Ecosystem Model

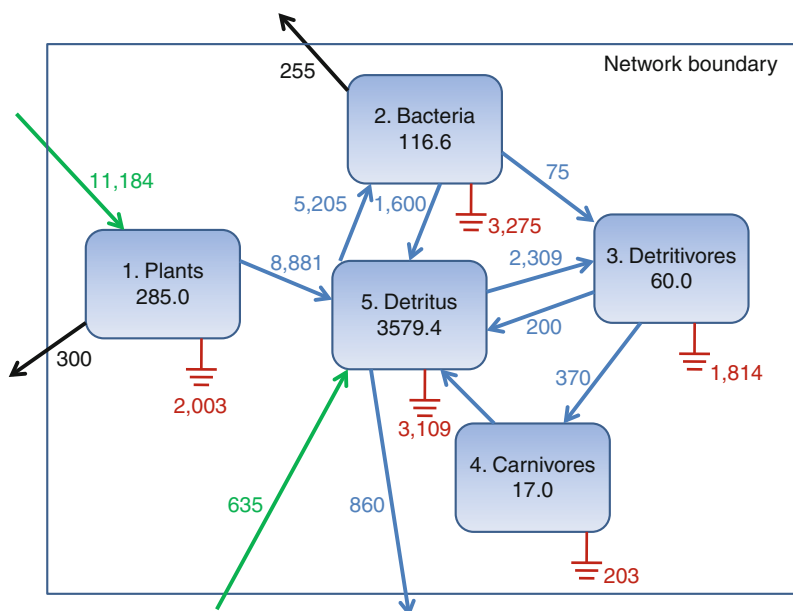
A classic example is the Cone Spring ecosystem model developed by Tilly [10]. In this model, there are five compartments representing: (1) plants, (2) bacteria, (3) detritivores, (4) carnivores, and (5) detritus (Fig. 3). There are 2 external inputs (to plants and detritus), 8 internal flows, and each compartment has boundary outflow representing metabolic or egestion losses. The internal flows from columns j to rows i are given by:

$$\mathbf{F} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 8881 & 0 & 1600 & 200 & 167 \\ 0 & 5205 & 0 & 0 & 0 \\ 0 & 2309 & 75 & 0 & 0 \\ 0 & 0 & 0 & 370 & 0 \end{bmatrix}$$



Ecosystem Flow Analysis. Figure 2

Schematic of flow partitioning for a central node i , in relation to other network compartments. Flow reaches the node directly across the boundary, $f^{(0)}$, by passing through other compartments before reaching i , $f^{(1)}$, and leaving i to cycle back again, $f^{(2)}$. Outflows symmetrically mirror these inputs



Ecosystem Flow Analysis. Figure 3

Network diagram of Cone Spring ecosystem energy flows [10]. All flows are in kcal/m²/year. Biomasses are in kcal/m². Green arrows are exogenous boundary inflows. Black arrows are exports of useable energy. Red ground symbols represent metabolic energy loss

Compartmental throughflows are: $T = [11\ 184, 11\ 484, 5\ 204, 2\ 384, 370]$ and $TST = 30\ 627$.

The nondimensional flow fractions are given by:

$$\mathbf{G} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.794 & 0 & 0.308 & 0.0840 & 0.451 \\ 0 & 0.453 & 0 & 0 & 0 \\ 0 & 0.201 & 0.014 & 0 & 0 \\ 0 & 0 & 0 & 0.155 & 0 \end{bmatrix}$$

And the integral flow matrix is:

$$\mathbf{N} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0.958 & 1.207 & 0.374 & 0.186 & 0.545 \\ 0.434 & 0.547 & 1.169 & 0.084 & 0.247 \\ 0.199 & 0.251 & 0.092 & 1.039 & 0.113 \\ 0.031 & 0.039 & 0.014 & 0.161 & 1.018 \end{bmatrix}$$

Looking at Table 2, it is seen that over 38% of the flow comes directly to a node from the first instant across a system boundary and 52% of the flow originates from one compartment and enters another compartment for the first time without cycling. Slightly over 9% of the total energy flow is material that has cycled by exiting and reentering the same compartmental node. In other words, about 3,000 kcal/m²/y of the total system throughflow is comprised of energy due to cyclic pathways which retain the energy in the system. This additional boost is important to the overall function of the Cone Spring ecosystem. A noticeable contribution of cycled flow is a common phenomenon in all ecosystems. Another way to demonstrate this importance of cycling, and the fundamental shift it has on how an ecosystem should be viewed, was given by Braner [11]. While investigating the same five-compartment Cone Spring model above, he showed

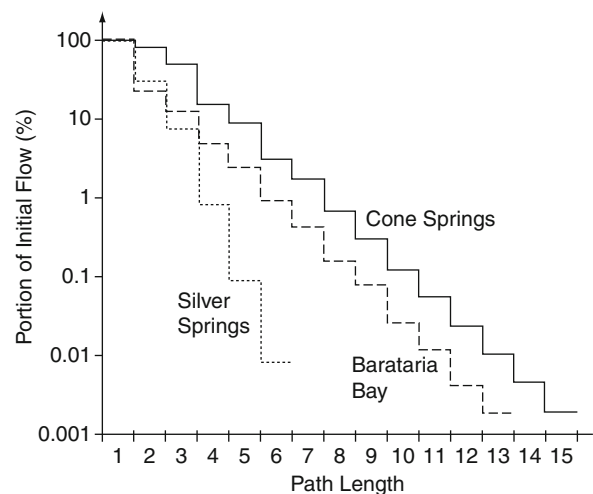
Ecosystem Flow Analysis. Table 2 Results of the flow analysis partitioning for Cone Spring ecosystem. Values represent the fraction of total system throughflow for each mode

Mode 0	0.386
Mode 1	0.522
Mode 2	0.092
Mode 3	0.522
Mode 4	0.386

that cyclic pathways identified by flow analysis reveal that the original boundary flow persists in the system much longer than obviously apparent. For contrast, in a five-compartment food chain model – a type often used, incorrectly, to represent an ecosystem – the longest path could only be four steps in length from $X1 \rightarrow X2 \rightarrow X3 \rightarrow X4 \rightarrow X5$. Real ecosystems have more complex structures with cycles. After those four steps, the original flow from compartment 1 would exit the system at compartment 5. According to Braner, more than 10% of the flow remains in the Cone Spring ecosystem after four steps. In fact, approximately 1% of the original flow remains after 9 steps and 0.001% is left after 15 steps (Fig. 4). A similar result is shown for two other ecosystems in the same figure. Therefore, the cycles, evident from flow analysis, play a very important role in the system having enough resource to function and provide ecosystem services.

Ecosystem Goal Functions

Flow analysis has another useful feature related to understanding ecosystem dynamics. Odum [12] proposed 24 different attributes which describe the ecosystem development, for aspects such as community



Ecosystem Flow Analysis. Figure 4

Portion of original energy that remains as path length increases (From Braner [11], reprinted with permission). There is still noticeable energy in the system after 10 steps

energetics, nutrient dynamics, and overall homeostasis. The attributes dealing with energetics, which change during the ecosystem development, have also been formulated as ecological goal functions – which describe observable macroscopic patterns over time. They are not strict goal functions in the sense of mathematical optimization models (neither is economic utility although it is used as such), but indicate the tendency for ecosystems to follow during development, for example, during succession from early r-selected species and r-selecting environments to late K-selected species and K-selecting environments. Some of the more common goal functions employed include: maximum power [13], maximum dissipation [14], maximum cycling [15], maximum residence time [16], minimum specific dissipation [17, 18], maximum energy [19], and maximum ascendancy [20] (see [9], for a detailed description of these). The idea is that the ecological network self-organizes itself in a way that leads to directional change in the property of these values. For example, maximum power, interpreted to mean the maximum throughflow in the network is given by: $TST = f^{(0)} + f^{(1)} + f^{(2)}$. Therefore, TST increases when there is more boundary flow (mode 0), more first passage flow (mode 1), or more cycled flow (mode 2). The mechanisms for this to increase practically relate to the system's ability to capture more boundary flow by increasing the uptake. Both first passage flow and cycled flow also depend on the second stage of growth exemplified by the structure of the network and the efficiency of flows along each connection. Similar rationale can be made for the other goal functions listed above, and in fact it has been shown that the goal functions are complementary and mutually reinforcing in that the realization of one generally promotes the others. Together they provide a holistic view of ecosystem development through different thermodynamic perspectives. Again, the value of this ecosystem knowledge is obvious for application to design and to manage human systems sustainably. If ecosystem services are required, then the inherent dynamics of the systems used should be better understood. Human activities in line with these directions will be supported by natural processes, those that do not will experience additional resistance and therefore additional cost and difficulty. Humans are better off working with nature than against it if possible.

Conclusions and Future Directions

Ecosystem flow analysis clearly shows that the distribution of energy flow in a network is not simple. Some significant fraction of the energy remains in the system and cycles before exiting the system. This insight was evident in R. Lindeman's [21] seminal work on Cedar Bog Lake in which he referred to his eight-compartment ecosystem as a "food-cycle." Unfortunately, he did not have the quantitative tools at his disposal, like flow analysis, and to simplify the calculations, proceeded to analyze the system according to two distinct "food-chains," although in reality they are linked and contain cycles. Further work in this area also neglected the presence and significance of food cycles until research in the mid-1970s (such as [22–25], and others) when network analysis techniques developed sufficiently to provide a holistic investigation of the ecosystem function. As stated above, this changes the way one must look at ecosystems, as processors and stores of energy flow. The energy does not pass quickly through but can remain and impact the system indirectly. The good news is that this flow which remains in the system is able to positively drive ecosystem processes and contribute to the overall sustainability of the system.

However, the lesson to take is that in the design of human systems, industrial processes are built sequentially, which have raw material → processing → product → disposal. There is little room for cycling and reuse. Remember, there are no trash cans in nature. Everything has a use and reuse. Efforts are now seen in industrial ecology promoting closed loop engineering and cradle-to-cradle considerations, but there is a long way to go, as evidenced by the massive amounts of raw material input and solid waste generated on a daily basis by human activity. Also, the flow analysis must include all parts and processes of the holistic integrated socio-ecological system. Future work is needed to continue to understand energy cycles in natural systems and furthermore, how to implement lessons from these into the design of socio-ecological systems. Ecosystem flow analysis clearly shows the input–output orientation flow resources have at their disposal for maintaining functional activity and can aid in sustainability science.

Bibliography

Primary Literature

1. Daly HE, Townsend KN (1993) *Valuing the earth: economics, ecology, ethics*. MIT Press, Cambridge, MA, p 399
2. Diamond J (2005) *Collapse: how societies choose to fail or succeed*. Penguin Press, New York, p 575
3. Fath BD, Jørgensen SE, Patten BC, Straškraba M (2004) Ecosystem growth and development. *Biosystems* 77:213–228
4. Hannon B (1973) The structure of ecosystems. *J Theor Biol* 41:535–546
5. Levine SH (1977) Exploitation interactions and structure of ecosystems. *J Theor Biol* 69:345–355
6. Fath BD, Scharler U, Ulanowicz RE, Hannon B (2007) Ecological network analysis: network construction. *Ecol Model* 208:49–55
7. Finn JT (1976) Measures of ecosystem structure and function derived from analysis of flows. *J Theor Biol* 56:363–380
8. Higashi M, Patten BC, Burns TP (1993) Network trophic dynamics: the modes of energy utilization in ecosystems. *Ecol Model* 66:1–42
9. Fath BD, Patten BC, Choi JS (2001) Complementarity of ecological goal functions. *J Theor Biol* 208:493–506
10. Tilly LJ (1968) The structure and dynamics of Cone Spring. *Ecol Monogr* 38:169–197
11. Braner M (1985) A simple random path method for the analysis of flow networks. *Ecol Model* 28:165–180
12. Odum EP (1969) The strategy of ecosystem development. *Science* 164:262–270
13. Lotka AJ (1922) Contribution to the energetics of evolution. *Proc Natl Acad Sci USA* 8:147–151
14. Schneider ED, Kay JJ (1994) Life as a manifestation of the second law of thermodynamics. *Math Comput Model* 19:25–48
15. Morowitz HJ (1968) Energy flow in biology; biological organization as a problem in thermal physics. Academic, New York
16. Cheslak EF, Lamarra VA (1981) The residence time of energy as a measure of ecological organization. In: Mitsch WJ, Bossermann RW, Klopatek JM (eds) *Energy and ecological modelling*. Elsevier, Amsterdam, pp 591–600
17. Onsager L (1931) Reciprocal relations in irreversible processes. *Phys Rev* 37:405–426
18. Prigogine I, Wiame JM (1946) *Biologie et Thermodynamique des Phenomenes Irreversible*. *Experientia* 11:451–453
19. Odum HT (1988) Self-organization, transformity, and information. *Science* 242:1132–1139
20. Ulanowicz RE (1997) *Ecology, the ascendent perspective*. Columbia University Press, New York
21. Lindeman RL (1942) The trophic dynamic aspect of ecology. *Ecology* 23:399–418
22. Pomeroy LR (1974) The ocean's food web, a changing paradigm. *Bioscience* 24:499–504
23. Finn JT (1976) Measures of ecosystem structure and function derived from analysis of flows. *J Theor Biol* 56:363–380
24. Patten BC (1978) Systems approach to the concept of environment. *Ohio J Sci* 78:206–222
25. Patten BC (1985) Energy cycling in the ecosystem. *Ecol Model* 28:1–71

Books and Reviews

- Burns TP (1989) Lindeman's contradiction and the trophic structure of ecosystems. *Ecology* 70:1355–1362
- Finn JT (1980) Flow analysis of models of the Hubbard Brook ecosystem. *Ecology* 61:562–571
- Latham LG (2006) Network flow analysis algorithms. *Ecol Model* 192:586–600
- Ulanowicz RE (1983) Identifying the structure of cycling in ecosystems. *Math Biosci* 65:219–237
- Whipple SJ (1999) Analysis of ecosystem structure and function: extended path and flow analysis of a steady-state oyster reef model. *Ecol Model* 114:251–274

Ecosystem Services

HEATHER TALLIS¹, ANNE GUERRY¹, GRETCHEN C. DAILY^{1,2}

¹The Natural Capital Project, Department of Biology and Woods Institute, Stanford University, Stanford, CA, USA

²Department of Biological Sciences, Stanford University, Stanford, CA, USA

Article Outline

Glossary
 Definition
 Introduction
 What are Ecosystem Services?
 History of the Concept of Ecosystem Services
 Incorporating Ecosystem Services into Decisions
 Future Directions
 Conclusions
 Bibliography

Glossary

Ecosystem services The wide array of benefits that ecosystems, and their biodiversity, confer on humanity.

Marginal value The economic value of the next incremental unit of something. In this context, marginal values are those associated with managing the next

small unit of an ecosystem in a particular way (e.g., preserving, rather than clearing, the next unit of forest). They can also be the partial contribution of natural capital to a final good that is produced with other inputs. For example, the marginal value of irrigation water for crop production is the value of the incremental crop yield that can be attributed to irrigation, rather than to labor, fertilizer, and other inputs.

Natural capital Here we focus on living, renewable forms of natural capital, which constitute a stock – of an ecosystem and the biota that makes it up – that generates a flow of ecosystem services. For example, a forest constitutes a stock that generates a flow of timber, carbon sequestration, water quality, biodiversity, serenity, and other benefits, depending upon how it is managed. (Fossil fuels and other minerals constitute nonliving natural capital, which is generally nonrenewable on time scales of interest to society.)

Definition

Ecosystem services are essential to sustaining and fulfilling human life, and yet their supply is seriously threatened by the intensification of human impacts on the environment. Over the past decade, efforts to value and protect ecosystem services have been promoted by many as the last best hope for making conservation mainstream – attractive and commonplace worldwide. In theory, if institutions recognize the values of nature, then we can greatly enhance investments in conservation and foster human well-being at the same time. In practice, scientific and policy communities have not yet developed the scientific basis or the policy and finance mechanisms for integrating natural capital into resource and land-use decisions on a large scale.

Introduction

This entry provides an overview of issues concerning the identification, biophysical and economic characterization, and safeguarding of ecosystem services. The concept of ecosystem services has a long written history, reaching back at least as far as Plato. We review this history, including a focus on recent advances such

as the Millennium Ecosystem Assessment, and advances in spatial modeling, economic valuation, and policy implementation. We provide examples of novel finance and policy mechanisms, including water funds, marine spatial planning, land-use planning and human development, and global policy efforts. We conclude with a discussion of the largest research and implementation challenges in this field, identifying the issues that will frame the future growth of the concept of ecosystem services.

What are Ecosystem Services?

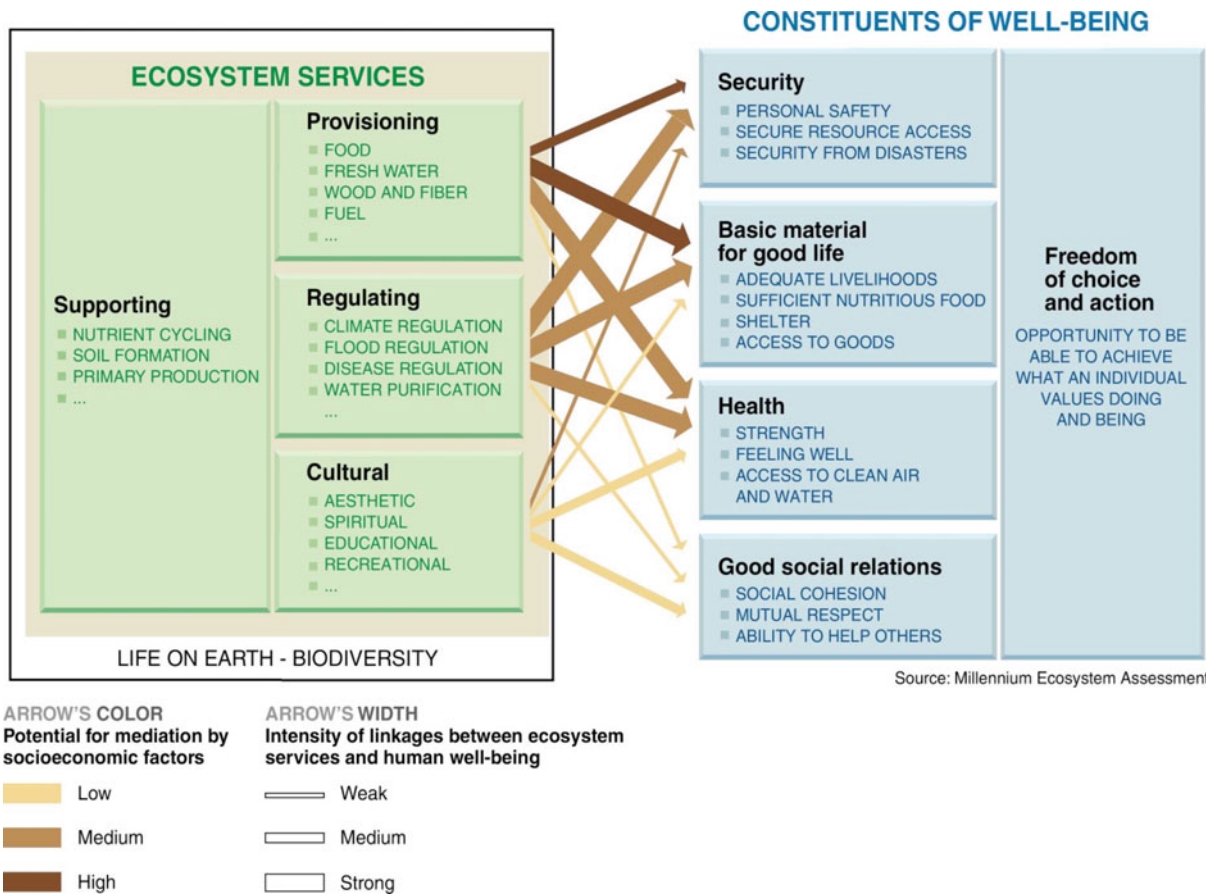
Definition and Classification

Ecosystem services are defined simply as the benefits that people obtain from ecosystems [38]. They sustain and fulfill human life and flow from many conditions and processes of ecosystems, and the species making them up [14]. The processes and features generating ecosystem services (ES) are so tightly interconnected that any classification is inherently somewhat arbitrary. The most widely used classification was developed through the Millennium Ecosystem Assessment (MA) and identifies four classes of ES based on their types of benefits to society:

1. *Provisioning services* including the production of goods such as food, water, timber, and fiber
2. *Regulating services* that stabilize climate, moderate risk of flooding and disease, and protect or enhance water quality
3. *Cultural services* that provide recreational, aesthetic, educational, community, and spiritual opportunities
4. *Supporting services* that underlie provision of the other three classes of benefits, including soil formation, photosynthesis, nutrient cycling, and the preservation of options (Fig. 1; [38]).

The classification of ES is still a topic of debate and several other classification approaches have been suggested [7, 19, 22, 67].

System- and scale-neutral, the ecosystem services framework applies equally to terrestrial, freshwater, and marine ecosystems and their processes, anywhere on the spectrum from relatively pristine to heavily managed conditions. Indeed, all ecosystems provide, to differing degrees, a set of ES. Human conversion of ecosystems from one type to another is often motivated



Ecosystem Services. Figure 1

Ecosystem service categories and their linkages to human well-being as described in the Millennium Ecosystem Assessment [38]

by a desire for a different set of ES, though consideration of the services of the two systems and their tradeoffs is often incomplete.

Ecosystem Services Across Systems

A wide range of ES is generated in the terrestrial realm, by croplands, natural and managed forests, grasslands, and wetlands. In each of these systems, for example, vegetation can protect and enhance soils, preventing their loss through erosion and improving fertility by retaining moisture and storing and recycling nutrients. Vegetation and soils together regulate the quantity, quality, and timing of water flows, thus moderating floods and droughts and providing cleaner, more reliable supplies [8].

Forests stand out as important in regulating water and carbon cycles [30]; in their strong influence on local, regional, and global climate [30, 49]; and because of the multiple, interacting threats to their future (e.g., [44, 66]). They also provide natural products for subsistence use or sale including timber, firewood, mushrooms, fruits and seeds, medicinal plants, rubber, cork, and bushmeat. Forest and woodland habitats harbor species that provide pollination and pest control to commercial or subsistence crops. Grassland and other dryland systems play these same critical roles in addition to supporting vast livestock populations [38]. Wetlands occupy a small fraction of Earth's surface, but dominate the landscape where they are concentrated and provide a wide array of water

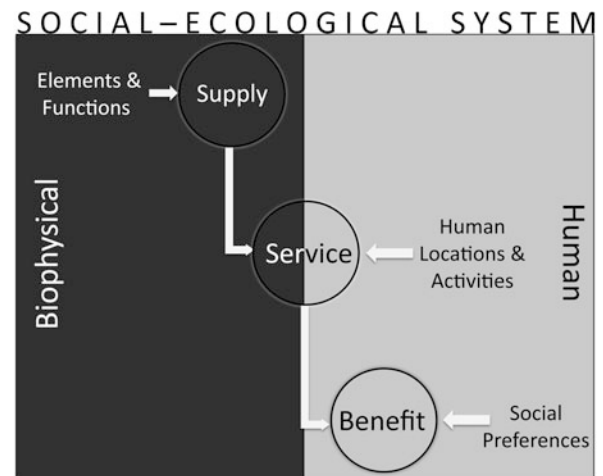
quality, flood mitigation, coastal protection, and biogeochemical services [38]. Each of these systems, however natural or managed, can provide habitat for biodiversity and opportunities for recreational activities, spiritual experiences, and creative, cultural expression.

Freshwater ecosystems provide a suite of highly visible and widely appreciated ES [53]. The freshwater regulated by terrestrial systems and the atmosphere is used for drinking, hydropower production, irrigation, household activities (washing, etc.), industrial purposes (cooling, manufacturing, etc.), and cultural experiences. People also gain large revenues and nutrition from freshwater fisheries and aquaculture. Less appreciated is the value of sediment transport and deposition in rivers that supply river reaches and downstream beaches with important sand and gravel resources. Wetlands and other aquatic vegetation can regulate flood waters and cycle nutrients, improving water quality. Finally, freshwater systems serve as pathways for human transportation and recreational or cultural activities.

Marine ecosystems also provide all four classes of ecosystem services described in the MA. Marine fisheries and aquaculture provide nutrition, feed for animals, livelihoods, and important recreational and cultural opportunities. Harvests of other species for food additives, cosmetics, and pharmaceuticals also support health, nutrition, and livelihoods. Marine biogenic habitats (such as coral reefs, oyster reefs, and kelp forests) regulate natural hazards including storm surges, and may play a critical role in helping coastal communities adapt to sea level rise. Marine systems also transform, detoxify, and sequester wastes. In addition, oceans are the center of the global water cycle; they hold 96.5% of the Earth's water [24] and are a primary driver of the atmosphere's temperature, moisture content, and stability [12]. Oceans are also key players in the global cycles of carbon, nitrogen, oxygen, phosphorus, sulfur, and other major elements [51] and are responsible for approximately 40% of global net primary productivity [39, 61]. Finally, coastal communities reap many benefits from coastal tourism (one the one of world's most profitable industries [65]), and numerous coastal communities define their very identities in relation to the sea and all it brings.

The Ecosystem Service Supply Chain

Ecosystem services flow to people along a supply chain from biophysical systems to people [63]. All services are generated by some function or element of a natural or managed system (Fig. 2). The full suite of these elements or functions can best be considered in three discrete steps: supply, service, and benefit. For example, consider protection from coastal storm surges. Many different types of coastal elements (e.g., coral reefs, mangroves, oyster beds, barrier islands) confer protection from storm surges by attenuating waves. The full set of locations of these coastal elements represents the supply of protection from storms. People do not receive storm surge protection from all of these locations, however, because some are far from human infrastructure and settlement. Both the distribution of human infrastructure and settlement, together with the location and condition of supply, are required to give a clear picture of how much "service" is actually delivered at a given time. Finally, the service delivered to human communities is often valued differently, depending on the context. For example, coastal protection services provided by nearshore habitats to easily



Ecosystem Services. Figure 2

Three measurement points for ecosystem services [63]. Supply metrics deal only with the biophysical system underpinning the service of interest. Service metrics include critical information linking supply to beneficiaries. Benefit metrics weigh the level of service based on people's preferences or social policy goals

accessible, popular, public beaches might be seen as more valuable, or providing greater benefit, than those to more remote sites.

History of the Concept of Ecosystem Services

It is primarily through disruption and loss that the nature and value of ecosystem services has been illuminated. For instance, deforestation has demonstrated the critical role of forests in the hydrological cycle – in particular, in mitigating floods, droughts, the erosive forces of wind and rain, and the silting of dams and irrigation canals. Release of toxic substances, whether accidental or deliberate, has revealed the nature and value of physical and chemical processes, governed in part by a diversity of microorganisms, that disperse and break down hazardous materials. Thinning of the stratospheric ozone layer sharpened awareness of the value of its service in screening out harmful ultraviolet radiation. And the loss of coastal wetlands has brought into relief their importance in regulating coastal hazards such as hurricanes and tsunamis.

Initial Development of the Ecosystem Services Concept

A cognizance of ecosystem services, expressed in terms of their loss, dates back at least to Plato and probably much earlier:

- What now remains of the formerly rich land is like the skeleton of a sick man with all the fat and soft earth having wasted away and only the bare framework remaining. Formerly, many of the mountains were arable. The plains that were full of rich soil are now marshes. Hills that were once covered with forests and produced abundant pasture now produce only food for bees. Once the land was enriched by yearly rains, which were not lost, as they are now, by flowing from the bare land into the sea. The soil was deep, it absorbed and kept the water..., and the water that soaked into the hills fed springs and running streams everywhere. Now the abandoned shrines at spots where formerly there were springs attest that our description of the land is true. (Plato)

Mooney and Ehrlich [40] trace modern concern for ecosystem services to George Perkins Marsh, a lawyer, politician, and scholar. Indeed, his 1864 book *Man and*

Nature describes a wide array of services, again, often expressed in terms of their loss. Remarking on the terrain of the former Roman Empire, he notes that it “is either deserted by civilized man and surrendered to hopeless desolation, or at least greatly reduced in both productiveness and population” (p. 9). He continues, describing the reduction of hydrological services: “Vast forests have disappeared from mountain spurs and ridges, the vegetable earth ... [is] washed away; meadows, once fertilized by irrigation, are waste and unproductive, because ... the springs that fed them dried up; rivers famous in history and song have shrunk to humble brooklets” (p. 9). He also draws connections between deforestation and climate: “With the disappearance of the forest, all is changed. At one season, the Earth parts with its warmth by radiation to an open sky – receives, at another, an immoderate heat from the unobstructed rays of the sun. Hence the climate becomes excessive, and the soil is alternately parched by the fervors of summer, and seared by the rigors of winter. Bleak winds sweep unresisted over its surface, drift away the snow that sheltered it from the frost, and dry up its scanty moisture” (p. 186). Finally, he even wrote of decomposition services: “The carnivorous, and often the herbivorous insects render an important service to man by consuming dead and decaying animal and vegetable matter, the decomposition of which would otherwise fill the air with effluvia noxious to health” (p. 95).

Other eloquent writers on the environment emerged following World War II, including Fairfield Osborn (*Our Plundered Planet*, 1948), William Vogt (*Road to Survival*, 1948), and Aldo Leopold (*A Sand County Almanac and Sketches from Here and There*, 1949). Each discusses ecosystem services without using the term explicitly. In *The Population Bomb* (1968), Paul Ehrlich describes anthropogenic disruption of ecosystems and the societal consequences of doing so, addressing the need to maintain important aspects of ecosystem functioning. Along these lines, the *Study of Critical Environmental Problems* (1970) presents a list of key “environmental services” that would decline with a decline in “ecosystem function.” This list was expanded upon by Holdren and Ehrlich [29]. Meanwhile, in the 1960s and 1970s, economists set out to measure “the value of services that natural areas provide” ([35], p. 12), with efforts focused on

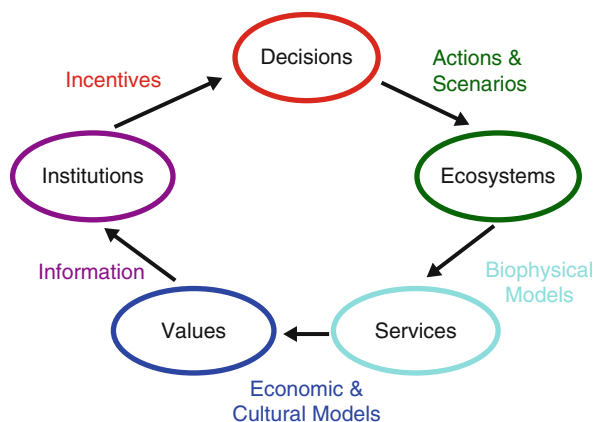
agricultural production [3], renewable resources [11, 34], nonrenewable resources [18], and environmental amenities [23].

By the early 1980s, efforts were initiated to investigate two questions: the extent to which ecosystem function (and the delivery of services) depends on biodiversity, and the extent to which technological substitutes could replace ecosystem services. The first question is addressed in chapter ► [Species Diversity Within and Among Ecosystems](#), this volume. The second question was tackled by Ehrlich and Mooney [21]. Work on these topics proliferated and, in 1997, a collective effort was made to synthesize the wealth of scientific information that had accumulated on the functioning of ecosystem services, with a preliminary exploration of their economic value, and of key issues meriting further work [14].

Recent Advances

Four major advances of the last decade have revitalized research on ecosystem services and brought them into the public eye. First, the MA represented a visionary and seminal step in global science – it was the first comprehensive global assessment of the status and trends of all of the world's major ecosystem services. It was requested by United Nations Secretary General Kofi Annan in 2000 and carried out between 2001 and 2005 with contributions from over 1,360 experts worldwide. The key finding of this assessment was that two thirds of the world's ecosystem services were declining [38]. This captured the attention of world leaders and emphasized the connections between human decisions and the natural environment that feed back to the human condition via changes in the flow of ecosystem services.

Work following the MA clarified this chain of connections (Fig. 3) [16]. Human decisions shape individuals' actions relating to the use of land, water, oceans, and other elements of natural capital. These actions often alter the state or functioning of ecosystems, which in turn provide altered flows of benefits (goods or services) to people. People express different values (monetary, cultural) associated with these altered streams of benefits and it is the expression of these values that leads to changes in institutions that guide decisions. The following three recent advances all concern the connections in this flow.



Ecosystem Services. Figure 3

A framework showing key elements for integrating ES into decision-making [16]. One could link any two ovals, in any direction, in different decision contexts

A suite of recent advances has greatly improved understanding of the links between ecosystem functions and processes and the provision of ecosystem services (Fig. 3). For some ecosystem services, we now better understand the key ecological system components that drive provision (e.g., [33]) and we can now measure (e.g., [56]) and model, with uncertainty, the impacts of land use and resource management decisions on a wider variety of ecosystem processes and associated services. Ecological science has also advanced spatially explicit modeling, which is essential for mapping ecosystem services and their flows to people (e.g., [10, 27, 43, 57]). Finally, we are starting to see patterns in how multiple ecosystem services and biodiversity change in relation to each other. Recent work has started characterizing bundles of ecosystem services, and exploring their synergies and trade-offs (e.g., [4, 6, 20, 42, 43]).

Further, economic valuation methods have been applied to the spatial provision of ecosystem services to estimate the monetary value of benefits and, in some cases, the distribution of those benefits to various segments of society [42, 45, 52, 64]. In addition, qualitative and quantitative methods from other social sciences have been applied to gain better understanding of the social and cultural importance of ecosystem services (e.g., [38]).

Lastly, experiments in payments for ecosystem services [47, 48, 69], in ecosystem-based management [2],

and in regional planning have begun, giving us opportunities to learn about how science can play a role in altering institutions, and how institutions alter decisions and the resulting flow of ecosystem services. The following section describes some of these efforts in more detail.

Incorporating Ecosystem Services into Decisions

Today, the urgent challenge is to move from theory to practical implementation of ES tools and approaches in resource decisions taken by individuals, communities, corporations, and governments. The framework in Fig. 3 connects the science of quantifying services with valuation and policy to devise payment schemes and management actions that take account of ecosystem services. This connection is expressed in the real world in a variety of ways across scales from local to global.

A great number and diversity of efforts to implement the ES framework have emerged worldwide over the past decade. Individually, most of these efforts are small and idiosyncratic. But collectively, they represent a powerful shift in the focus of conservation organizations and governments (primarily) toward a more inclusive, integrated, and effective set of strategies [15]. Taken together, these efforts span the globe and target a full suite of ecosystem services, including principally forest-generated services of carbon sequestration, water supply, flood control, biodiversity conservation, and enhancement of scenic beauty (and associated recreation/tourism values) [26, 62].

Many local or regional ES efforts focus on a single service that stands out as sufficiently important, from economic and political perspectives, to overcome the activation energy required to protect it. Under the institutional umbrella created for the focal service, it is possible that other services may be at least partially protected. Beginning in the late 1990s, larger-scale investment in natural capital for water flow regulation in China – and for a broad suite of ES in Costa Rica – set pioneering examples that are now being adapted elsewhere and scaled up.

Next, we briefly describe some contrasting models of success, at different scales and in different kinds of social-ecological systems. In each case, there is an acute or looming crisis, innovative leadership, and pursuit of dual goals: improving both human and ES condition.

Local Scale: Water Funds

New York City made one of the first and most famous investments in ecosystem service provision in recent history, in the mid-1990s. The city invested ca. USD1.5 billion in a variety of watershed protection activities to improve drinking water quality for 10 million users rather than spending the estimated USD6–8 billion needed (excluding annual operating and maintenance costs) for building a new filtration plant. This seminal example is widely cited as evidence of the business case for investing in natural capital instead of built capital [15]. Yet the effort remains very much an experiment in the science and policy of investing in natural capital, and one on which there is international focus.

Globally, watersheds are now emerging as the target of a range of creative policy and finance mechanisms that link beneficiaries to suppliers through a payment system. In these “water funds,” water users voluntarily pay into a pool that is collectively managed by contributors and invested in watershed management improvements. The Nature Conservancy (TNC) has now established more than ten water funds in Latin America, has plans to create 22 more by 2015 [25], and is exploring the possibility of establishing some of the first funds in Africa.

Agua por la Vida y la Sostenibilidad, one of the recently established water funds, demonstrates the diversity of water users that are becoming engaged in these funds and the kinds of watershed management changes these funds motivate. Formally established in the Cauca Valley, Colombia in 2009, this water fund is supported by the region’s sugarcane grower’s association (PROCAÑA), the sugar producers’ association (ASOCAÑA), 11 local watershed management groups, TNC and a Colombian peace and justice nongovernment organization (Vallenpaz). Each member of the water fund voluntarily pays a self-determined amount into the fund that is then jointly managed by the members to improve landscape management in 11 watersheds covering over 3,900 km².

Members in this fund have currently committed to contributing USD10 million over 5 years to be invested in five kinds of management changes: protection of native vegetation, restoration of denuded lands, enrichment of degraded forests, fencing of rangelands, and implementation of silvopastoral practices.

The fund is starting a monitoring program that will ensure that these investments lead to measurable improvements in water quality for approximately one million water users downstream and significant improvements in terrestrial and freshwater biodiversity.

Local Scale: Coastal and Marine Spatial Planning

People commonly think of oceans as relatively featureless expanses that defy the drawing of lines on maps. However, recent political and scientific advances have highlighted the need for a comprehensive approach to planning marine and coastal uses and the need for practical tools to make this more comprehensive approach a reality on the ground and in the water. In a marine spatial plan, a wide range of uses of the marine environment are put on one map. But an understanding of how such plans are likely to yield changes in the delivery of the broad range of services people receive from the system has, until recently, remained elusive.

Along the west coast of Vancouver Island Canada, multiple, often competing interests are struggling to define the future character of the place. Existing extractive, industrial, and commercial uses; traditional First Nations subsistence and ceremonial uses; recreation and tourism; and emerging ocean uses such as the extraction of wave energy are all in the mix. The West Coast Aquatic Management Board (WCA) is charged with creating a marine spatial plan for the region. WCA is a public-private partnership with participation from four levels of government (Federal, Provincial, local, and First Nations), and diverse stakeholders. Ultimately, WCA's vision is to manage resources for the benefit of current and future generations of people and nonhuman species and communities.

Some key pillars of the partnership's strategy are to: use a precautionary, ecosystem-based approach to protect, maintain, and restore marine and coastal resources; respect and protect First Nations' food, social and ceremonial requirements and treaty obligations; integrate expertise and knowledge from First Nations, local, scientific, and other sources; ensure broad participation in the planning process; and foster initiatives that maintain or enhance opportunities for coastal communities to benefit from local resources, while achieving sustainable social, cultural, and economic benefits for the region. WCA has partnered with

the Natural Capital Project to explore how alternative spatial plans might affect a wide range of ES and to provide information about trade-offs among ES.

Key considerations for WCA and their stakeholders include balancing important industrial and commercial activities (such as shipping, mining, logging, aquaculture, and fisheries), increased development of tourism and recreation, renewable energy generation, and a strong cultural desire for sustaining the remote, wild feeling of the place. WCA is exploring the suitability of alternative regions for these different activities. For example, maps of coastal vulnerability to erosion and flooding from storm surge are helping to direct coastal development permits to low-risk areas. Similar maps of the value of captured wave energy are being overlaid with existing ocean uses (e.g., fishing and recreational activities) to highlight regions of high wave energy value, where wave energy generation facilities might be constructed while having minimal impacts on other activities. Examinations of trade-offs among aquaculture (finfish, shellfish), wild salmon fisheries, recreation (e.g., kayaking, whale watching, and diving), coastal development (on the coast, as well as floathomes), and habitat and water quality are underway.

The general framework of ES and ES modeling, in particular, is helping to articulate connections between human activities that are often considered in isolation, to align diverse stakeholders around common goals, and to make implicit decisions explicit. ES modeling results have informed early iterations of the marine spatial plan and will inform the creation of the final plan in 2012.

National Scale: Land-Use Planning and Human Development in China

The ecosystem service investments being made in China today are impressive in their goals, scale, duration, and innovation. Following massive droughts and flooding in 1997–1998, China implemented several national forestry and conservation initiatives, into which investments exceeded 700 billion yuan (ca. USD100 billion) over 2000–2010 [37, 70]. The larger and older of these initiatives are being rigorously evaluated to determine their biophysical and socioeconomic impacts, to improve their design and efficacy.

These initiatives have dual goals: to secure critical natural capital through targeted investments across landscapes and regions, and to alleviate poverty through targeted wealth transfers from coastal provinces to inland regions where many ES originate. The Chinese government aims to reduce the loss of soil, improve water retention, reduce desertification, and generally protect biodiversity and ecosystems in the west of the country for flood control, hydropower production efficiency, irrigation supply, more productive agriculture, and ecotourism. In addition, it wants to change the economic structure in rural areas to increase local household income while simultaneously making local households' patterns of land utilization and agricultural production more sustainable [36, 37].

The initiatives include two national PES programs, the Natural Forest Conservation Program (NFCP) and the Sloping Land Conversion Program (SLCP), established in 1998 and 1999 respectively. Implementation was tested in a few provinces, and then rapidly scaled to the whole country. Evaluation of the programs shows significant achievement of the biophysical goals, with remarkably rapid land conversion in the desired directions. For example, by the end of 2006, the SLCP had converted ca. 9 million hectares of cropland into forest/grassland and had afforested ca. 12 million hectares of barren land. Village-level field measurements have shown not only that the payments for ES have altered land use patterns, but in turn soil erosion has been decreased in some areas by as much as 68% [9].

Overall social impacts of the programs are mixed. In some places, payment levels and types are leading to improvements in economic measures of well-being, whereas in others payments were not sufficient to compensate for loss of income from shifting livelihoods [37]. In addition, in some places where participation in the SLCP has significant positive impacts upon household income, it has not yet transferred labor toward non-farming activities as the government wished [36]. Payments are now being adjusted to improve success in achieving goals of poverty alleviation and growth of new economic sectors in rural areas.

China is also now establishing a new network of Ecosystem Function Conservation Areas (EFCAs), specifically for ES provision. Their exact delineation is now being determined through quantitative ecosystem

service mapping and valuation. They are expected to span ca. 25% of the country.

The current and potential future impacts of ES investments in China are enormous, certainly within the country – and also globally, in the form of enhanced carbon sequestration and reduced dust export, and perhaps most importantly in lessons on making the investments needed in natural capital and human well-being everywhere.

International Scale: Global Policy and Research Efforts

As described above, the MA was the first major effort to establish ES in the international policy arena. Activities stemming from that effort are now aimed at bringing countries together in making tangible commitments to safeguard ES (e.g., 2020 targets for the Convention on Biodiversity) and to assess national and international progress toward those commitments (e.g., through Group on Earth Observations Biodiversity Observation Network (GEO BON) and the Programme on Ecosystem Change and Society (PECS), which synthesize knowledge for the International Platform for Biodiversity and Ecosystem Services (IPBES), formally established in 2010). Several new international research efforts aim to feed into these international processes, including the Natural Capital Project, The Resilience Alliance, and the Stockholm Resilience Centre. Other entities are focused on establishing and tracking ES markets, as a mechanism for bringing larger attention to ES benefits to society (e.g., The Katoomba Group and The Ecosystem Marketplace, both initiated by Forest Trends). As an example of many burgeoning international efforts, we describe in greater detail the Natural Capital Project.

The Natural Capital Project (NatCap) (www.naturalcapitalproject.org) is an international partnership working to align economic forces with conservation, by developing tools that make incorporating natural capital into decisions easy and replicable; by demonstrating the power of these tools in important, contrasting places; and by engaging leaders globally. NatCap is developing InVEST, a family of tools for Integrated Valuation of Ecosystem Services and Tradeoffs.

InVEST helps decision makers visualize the impacts of potential policies by modeling and mapping the

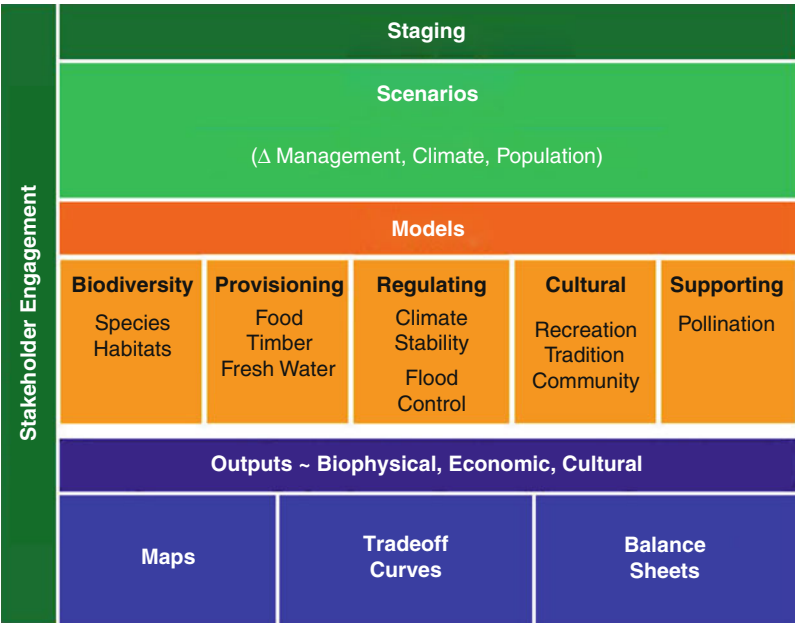
delivery, distribution, and economic value of ES under alternative scenarios (for more information, see [30]). The outputs identify trade-offs and compatibilities between environmental, economic, and social benefits. InVEST is designed for use as part of an active decision-making process (Fig. 4) and can be applied at local, regional, or global scales. The first phase of the approach involves working with stakeholders to identify critical management decisions and to develop scenarios that project how the provision of services might change in response to those decisions as well as to changing climate, population, etc. Based on these scenarios, a modular set of models quantifies and maps ES. The outputs of these models provide decision makers with information about costs, benefits, trade-offs, and synergies of alternative investments in ES provision.

NatCap is using InVEST in major natural resource decisions in diverse contexts around the world, including in the three examples given above (water funds, coastal and marine spatial planning, and land-

use planning and human development in China). The aim is to demonstrate the power of these approaches and to learn how to replicate and scale up models of success. The Project is engaged in a suite of international efforts, including GEO BON and IPBES, to offer a common, unifying platform for regional and national efforts that are spawned by these initiatives.

Future Directions

With the rapid rate of development of ES mapping, from the biophysical and economic modeling through to policy application in diverse socioeconomic contexts, it is likely that great advances will be made in coming years. What we report here is only a beginning. There are key arenas in which further learning is crucial to understand what drives variation in the provision of ES, how they percolate through various arms of society, and how social reaction leads to sustainable or unsustainable change in ES provision.



Ecosystem Services. Figure 4

An iterative process for integrating ecosystem services into decisions. The process begins with stakeholder engagement around impending decisions, with a focus on realistic, alternative scenarios for the future. The modeling is shaped by stakeholders, and typically focused on the services and scenarios deemed most important. Outputs are displayed in accordance with stakeholder preferences, in the form of maps, trade-off curves, and/or balance sheets. These can be expressed in biophysical (e.g., tons of carbon), economic (e.g., dollars), or cultural (e.g., visitor-days) terms

Relating Ecosystem Services and Human Health

The relationships between biophysical attributes of ecosystems and human health are complex [41]. Destruction of natural ecosystems can at times improve aspects of public health. Draining swamps, for example, can reduce habitat for the mosquito vector that transmits the parasite that causes malaria. On the other hand, destruction of other systems can have sharp negative consequences for human health. There is emerging evidence that loss of tropical forests, for example, leads to an increase in transmission of malaria [1, 50]. Similarly, fragmentation of, and biodiversity loss from, eastern North American forests is associated with an increase in lyme disease [32].

Natural and managed ecosystems provide many services that sustain human health, through provision of human nutrition (especially of protein and micronutrients); purification and regulation of drinking water; regulation of air quality; regulation of vector-borne disease; and psychological benefits. There is a great need for research illuminating the links between biodiversity, ecosystem conditions and processes, and human health.

Trade-offs and Synergies

The relationship between ES and biodiversity and among different ES varies with socio-ecological context. In some cases, clear trade-offs and synergies among services have been defined in specific contexts [54], but there is still much to be learned about what determines the nature of these relationships. Advancing this knowledge is essential because policies addressing management change can only be successful if management controls ES relationships. If policies are established to align multiple ecosystem services, but biophysical conditions in the system lead to innate trade-offs among services, management changes are bound to fail in delivering the desired improvements to social benefits.

Distributional Effects

Much of the science of mapping ES has focused on identifying where ES are generated and where they are delivered. However, less work has focused on identifying to whom ES actually flow. This connection is essential if policies addressing ES delivery are to be equitable

and either improve the well-being of the poor or avoid unintended distributional consequences. Past work in this arena has focused on overlaying maps of ES provision with an array of poverty indicators (e.g., [68]). Missing from this spatial analysis is information on access to and ability to control the delivery of ES. In many cases (e.g., for services such as clean drinking water, hydropower production, agriculture, water for irrigation, wave power generation), the actual delivery of services to specific people is affected by the location of infrastructure or institutions regulating access to resources. New science is needed that allows the ready mapping of these connections and the prediction of how they will change under future conditions.

Dynamic Effects: Shocks and Uncertainty

Dynamic changes, such as in climate and in the nitrogen cycle – as well as changes arising through economic development and evolving human preferences over time – are very important. The possibility of feedbacks within ecosystems, and between ES and human behavior, is a key area for further development. Feedback effects can give rise to thresholds and rapid changes in systems that can fundamentally alter system outcomes [60]. The ability to incorporate shocks and the possibility of surprises is another area where further development is needed. Fires, droughts, and disease all can have major influences on ecosystems and affect the services produced. Changes in economic conditions or fads in human behavior can similarly cause major changes in systems (e.g., financial crises). The occurrence of each of these and other potential disturbances is difficult to predict but virtually certain to come about. Understanding their likely impacts on ecological and social systems will help us prepare for them.

Valuation in Monetary and Non-monetary Terms for Decision-Making

Monetary valuation of ES is not nearly as prevalent as sometimes assumed. More typically, real-world applications of the ES framework rely on biophysical values to inform policy design, such as measures of water quality or flood risk.

Value is not always easily characterized or fully captured in monetary terms, so it is important to characterize value in multiple dimensions, including

health, livelihood support, cultural significance, etc. (e.g., [17]). This will help ensure that valuation and broader decision-making approaches are inclusive of the range of benefits and people concerned [28]. Interdisciplinary efforts are presently underway to create a conceptual framework that is useful both in theory and in practice for a broad suite of cultural ES.

Institutional Design

However ES are measured, there is a need for political and social science research to design institutions and policy mechanisms that better capture externalities. Efforts such as national accounts are blossoming now, but it is unclear how they will evolve and how successful governments will be at incorporating natural capital into national measures of wealth. There is great work to be done in determining the merits and limitations of alternative policy and finance mechanisms, in different economic, governance, and other social contexts (e.g., [5, 46, 58, 59]). There is also great work to be done in developing institutions that achieve representation and participation by stakeholders as part of adaptive governance systems (e.g., [13, 55]).

Conclusions

Ecosystem services have had a relatively long history through indirect recognition of the importance of nature for the persistence of the human endeavor. There are scientific challenges for biogeochemists, hydrologists, ecologists, economists, anthropologists, and other social scientists to understand how human actions affect ecosystems, the provision of ES, and the value of those services. At least as demanding are the social and political questions associated with incorporating this understanding into decision-making. There is also a need to design effective and enduring institutions to manage, monitor, and provide incentives that reflect the social values of ecosystem services. Information is becoming more readily available for individuals, corporate managers, and government officials who make decisions affecting ecosystems and the services to consider a more complete set of costs and benefits associated with their choices. We are likely to see continuing growth in our scientific ability to measure and predict changes in ES, our ability to design policies and institutions that accurately represent these changes

and in turn, the ability of the environment to continue providing the many benefits society needs to prosper.

Bibliography

Primary Literature

1. Afrane YA, Little TJ, Lawson BW, Githeko AK, Yan G (2008) Deforestation and vectorial capacity of *Anopheles gambiae* Giles mosquito in malaria transmission, Kenya. *Emer Infect Dis* 14:1533–1538
2. Barbier E, Koch E, Silliman B et al (2008) Coastal ecosystem-based management with non-linear ecological functions and values. *Science* 319:321–323
3. Beattie B, Taylor CR (1985) *The economics of production*. Wiley, New York
4. Bennett E, Peterson G, Gordon L (2009) Understanding relationships among multiple ecosystem services. *Ecol Lett* 12:1394–1404
5. Berkes F, Colding J, Folke C (2003) *Navigating social-ecological systems*. Cambridge University Press, Cambridge
6. Boody G, Vondracek B, Andow D et al (2005) Multifunctional agriculture in the United States. *Bioscience* 55:27–38
7. Boyd JW, Banzhaf HS (2005) Ecosystem services and government accountability: the need for a new way of judging Nature's value. *Resources* 158:16–19, Summer
8. Brauman KA, Daily GC, Duarte TK, Mooney HA (2007) The nature and value of ecosystem services: an overview highlighting hydrologic services. *Annu Rev Environ Resour* 32:67–98
9. Cao S, Zhong B, Yue H, Zeng H, Zeng J (2009) Development and testing of a sustainable environmental restoration policy on eradicating the poverty trap in China's Changting County. *Proc Natl Acad Sci* 106:10712–10716
10. Chan K, Shaw R, Cameron D et al (2006) Conservation planning for ecosystem services. *Public Libr Sci Biol* 4:2138–2152
11. Clark C (1990) *Mathematical bioeconomics*, 2nd edn. Wiley, New York
12. Colling A (2001) *Ocean circulation*. Butterworth Heineman/The Open University, Milton Keynes
13. Cowling R, Egoh B, Knight AT et al (2008) An operational model for mainstreaming ecosystem services for implementation. *Proc Natl Acad Sci* 105:9483–9488
14. Daily GC (ed) (1997) *Nature's services*. Island Press, Washington, DC
15. Daily GC, Ellison K (2002) *The new economy of nature*. Island Press, Washington, DC
16. Daily GC, Polasky S, Goldstein J et al (2009) Ecosystem services in decision-making: time to deliver. *Front Ecol Environ* 7:21–28
17. Dasgupta P (2001) *Human well-being and the natural environment*. Oxford University Press, Oxford
18. Dasgupta P, Heal G (1979) *Economic theory and exhaustible resources*. Cambridge University Press, Cambridge, UK
19. De Groot D, Wilson MA, Boumans RMJ (2002) A typology for the classification, description and valuation of ecosystem functions, goods and services. *Ecol Econ* 41:393–408

20. Egoh B, Reyers B, Rouget M et al (2008) Mapping ecosystem services for planning and management. *Agric Ecosyst Environ* 127:135–140
21. Ehrlich PR, Mooney HA (1983) Extinction, substitution, and ecosystem services. *BioScience* 33:248–254
22. Fisher B, Turner RK, Morling P (2008) Defining and classifying ecosystem services for decision making. *Ecol Econ* 68:643–653
23. Freeman AM III (1993) The measurement of environmental and resource values. Resources for the Future, Washington, DC
24. Gleick P (1996) Water resources. In: Schneider S (ed) *Encyclopedia of climate and weather*. Oxford University Press, New York, pp 817–823
25. Goldman R, Benítez S, Calvache A, Ramos A (2010) Water funds: protecting watersheds for nature and people. The Nature Conservancy, Arlington
26. Goldman RL, Tallis H, Kareiva P, Daily GC (2008) Field evidence that ecosystem service projects support biodiversity and diversify options. *Proc Natl Acad Sci* 105:9445–9448
27. Harrison J, Bouwman A, Mayorga E, Seitzinger S (2010) Magnitudes and sources of dissolved inorganic phosphorus inputs to surface fresh waters and the coastal zone: a new global model. *Glob Biogeochem Cycles* 24:271–279
28. Heal G (2000) *Nature and the marketplace: capturing the value of ecosystem services*. Island Press, Washington, DC
29. Holdren J, Ehrlich P (1974) Human population and the global environment. *Am Sci* 62:282–292
30. Jackson RB, Randerson JT, Canadell JG, Anderson RG, Avissar R, Baldocchi DD, Bonan GB, Caldeira K, Diffenbaugh NS, Field CB, Hungate BA, Jobbágy EG, Kueppers LM, Noss MD, Pataki DE (2008) Protecting climate with forests. *Environ Res Lett* 3:1–5
31. Kareiva PK, Tallis H, Ricketts TH, Daily GC, Polasky S (eds) (2011) *Natural capital: theory & practice of mapping ecosystem services*. Oxford University Press, Oxford
32. Keesing F et al (2010) Impacts of biodiversity on the emergence and transmission of infectious diseases. *Nature* 468:647–652
33. Kremen C (2005) Managing for ecosystem services: what do we need to know about their ecology? *Ecol Lett* 8:468–479
34. Krutilla JV (1967) Conservation reconsidered. *Am Econ Rev* 57:777–786
35. Krutilla JV, Fisher AC (1975) *The economics of natural environments: studies in the valuation of commodity and amenity resources*. Johns Hopkins University Press, Baltimore
36. Li J, Feldman M, Li S, Daily GC (2011) Rural household income and inequality under payment for ecosystem services: the Sloping Land Conversion Program in Western China. *Proc Natl Acad Sci* 108:7721–7725
37. Liu J, Li S, Ouyang Z et al (2008) Ecological and socioeconomic effects of China's policies for ecosystem services. *Proc Natl Acad Sci* 105:9489–9494
38. MA (Millennium Ecosystem Assessment) (2005) *Ecosystems and human well-being: the assessment series (four volumes and summary)*. Island Press, Washington, DC
39. Melillo JM, McGuire AD, Kicklighter DW, Moore B, Vorosmarty CJ, Schloss AL (1993) Global climate-change and terrestrial net primary production. *Nature* 363:234–240
40. Mooney HA, Ehrlich PR (1997) Ecosystem services: a fragmentary history. In: Daily GC (ed) *Nature's services*. Island Press, Washington, DC, pp 11–19
41. Myers S, Patz J (2009) Emerging threats to human health from global environmental change. *Annu Rev Environ Resour* 34:223–252
42. Naidoo R, Ricketts TH (2006) Mapping the economic costs and benefits of conservation. *PLoS Biol* 4:e360
43. Nelson E, Mendoza G, Regetz J et al (2009) Modeling multiple ecosystem services, biodiversity conservation, commodity production, and tradeoffs at landscape scales. *Front Ecol Environ* 7:4–11
44. Nepstad D, Carvalho G, Barros AC, Alencar A, Capobianco JP, Bishop J, Moutinho P, Lefebvre P, Silva UL, Prins E (2001) Road paving, fire regime feedbacks, and the future of Amazon forests. *For Ecol Manag* 154:395–407
45. NRC (National Research Council) (2005) *Valuing ecosystem services: toward better environmental decision-making*. National Academies, Washington, DC
46. Ostrom E (2005) *Understanding institutional diversity*. Princeton University Press, Princeton
47. Pagiola S, Arcenas A, Platais G (2005) Can payments for environmental services help reduce poverty? *World Dev* 33:237–253
48. Pagiola S, Bishop J, Landell-Mills N (2002) *Selling forest environmental services*. Earthscan, London
49. Parry ML, Canziani OF, Palutikof JP, van de Linden PJ, Hanson CE (eds) (2007) *Contribution of working group II to the fourth assessment report of the intergovernmental panel on climate change*. Cambridge University Press, Cambridge, UK
50. Patz JA, Olson SH (2006) Malaria risk and temperature: influences from global climate change and local land use practices. *Proc Natl Acad Sci* 103:5635–5636
51. Peterson C, Lubchenco J (1997) Marine ecosystem services. In: Daily G (ed) *Nature's services*. Island Press, Washington, DC
52. Polasky S, Nelson E, Pennington D, Johnson KA (2010) The impact of land-use change on ecosystem services, biodiversity and returns to landowners: a case study in the state of Minnesota. *Environ Resour Econ* 48:219–242
53. Postel S, Carpenter S (1997) Freshwater ecosystem services. In: Daily GC (ed) *Nature's services*. Island Press, Washington, DC, pp 195–214
54. Raudsepp-Hearne C, Peterson GC, Bennett EM (2010) Ecosystem service bundles for analyzing tradeoffs in diverse landscapes. *Proc Natl Acad Sci* 107:5242–5247
55. Rickenbach MG, Reed AS (2002) Cross-boundary cooperation in a watershed context: the sentiments of private forest landowners. *Environ Manage* 30:584–594
56. Ricketts T, Regetz J, Steffan-Dewenter I et al (2008) Landscape effects on crop pollination services: are there general patterns? *Ecol Lett* 11:499–515
57. Rokityanskiy D, Benítez P, Kraxner F et al (2007) Geographically explicit global modeling of land-use change, carbon sequestration, and biomass supply. *Technol Forecasting Soc Chang* 74:1057–1082

58. Ruhl JB, Kraft SE, Lant CL (2007) The law and policy of ecosystem services. Island Press, Washington, DC
59. Salzman J (2005) Creating markets for ecosystem services: notes from the field. *NYU Law Rev* 80:870–961
60. Scheffer M, Carpenter SR, Foley J et al (2001) Catastrophic shifts in ecosystems. *Nature* 413:591–596
61. Schlesinger W (1991) Biogeochemistry: an analysis of global change. Academic, San Diego
62. Tallis H, Goldman R, Uhl M, Brosi B (2009) Integrating conservation and development in the field: implementing ecosystem service projects. *Front Ecol Environ* 7:12–20
63. Tallis H, Lester SE, Ruckelshaus M, Plummer M, McLeod K, Guerry A, Andelman S, Caldwell MR, Conte M, Copps S, Fox D, Fujita R, Gaines SD, Gelfenbaum G, Gold B, Kareiva P, Kim C, Lee K, Papenfus M, Redman S, Silliman B, Wainger L, White C (2012) New metrics for managing and sustaining the ocean's bounty. *Mar Policy* 36:303–306
64. Troy A, Wilson M (2006) Mapping ecosystem services: practical challenges and opportunities in linking GIS and value transfer. *Ecol Econ* 60:435–449
65. United Nations Environment Programme (2006) Marine and coastal ecosystems and human well-being: a synthesis report based on the findings of the millennium ecosystem assessment. UN Environment Programme, Nairobi
66. van Mantgem PJ, Stephenson NL, Byrne JC, Daniels LD, Franklin JF, Fulé PZ, Harmon ME, Larson AJ, Smith JM, Taylor AH, Veblen TT (2009) Widespread increase of tree mortality rates in the Western United States. *Science* 323:521–524
67. Wallace K (2007) Classification of ecosystem services: problems and solutions. *Biol Conserv* 139:235–246
68. World Resources Institute (2007) Kenya poverty mapping
69. Wunder S, Engel S, Pagiola S (2008) Taking stock: a comparative analysis of payment for environmental services programs in developed and developing countries. *Ecol Econ* 65:834–852
70. Zhang P, Shao G, Zhao G et al (2000) China's Forest policy for the 21st century. *Science* 288:2135–2136

Ecosystems and Spatial Patterns

PATRICK M. A. JAMES¹, MARIE-JOSÉE FORTIN²

¹Department of Biological Sciences, University of Alberta, Edmonton, AB, Canada

²Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON, Canada

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Sources of Heterogeneity

Spatial Analyses
 Spatial Analyses within an Ecosystem
 Spatial Analyses Among Ecosystem
 Future Directions
 Acknowledgments
 Bibliography

Glossary

Disturbance A spatial process or event that reverts forest vegetation to early successional stages typically altering forest structure and composition.

Ecotone A region of interface between two communities, ecosystems, or biogeographic regions.

Legacy A persisting spatial feature or pattern that was generated by a historical disturbance. Legacies can constrain the spatial dynamics of contemporary disturbances.

Multi-scale analysis A method of spatial analysis that looks at the relative contributions of different scales of spatial pattern to a single observed spatial pattern.

Pattern A repeatable and identifiable feature in a spatial context.

Scale An attribute of a spatial process or data used to represent that process that describes its spatial dimensions. Scale includes elements of grain, extent, and thematic resolution.

Spatial autocorrelation The degree of correlation of a variable and itself as a function of the spatial distances among sample points.

Stationarity A feature of a spatial process in which the mean and variance of a process is consistent across the extent of a study area.

Variography A geostatistical modeling tool for describing spatial variance and semivariance as a function of spatial distance among pairs of points.

Definition of the Subject

Ecological processes such as forest disturbances act on ecosystems at multiple spatial and temporal scales to generate complex spatial patterns. These patterns in turn influence ecosystem dynamics and have important consequences for ecosystem sustainability. Analysis of ecosystem spatial structure is a first step toward understanding these dynamics and the uncertain interactions among processes. There are many spatial

statistics available to describe and test spatial pattern within ecosystems and to infer the character of the processes that generated them. Indeed, improving understanding of the processes that create spatial pattern is a central objective of spatial pattern analysis. In addition to standard tests of spatial autocorrelation and patch structure, methods for multi-scale decomposition of spatial data and identification of stationarity are necessary to determine the key spatial scales at which the processes operate and affect ecosystems and to identify meaningful spatial subunits within larger contexts. Finally, tools for identifying ecosystem boundaries are also important to monitor boundary movement and changes in local ecosystem characteristics through time.

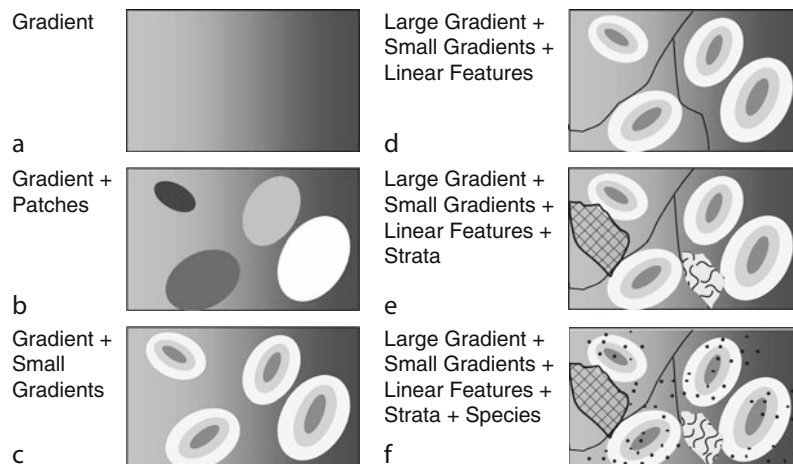
Introduction

Spatial Patterns in Ecosystems

The spatial structure of ecological systems is important to examine and understand as spatial structure mediates the flows of individuals, materials, and information through space and time [1]. These flows bear on the probabilities of occurrence and persistence of floral and faunal populations which determine local and regional biological diversity as well as ecosystem

functioning [2]. Interruptions and alterations of such flows within and among ecosystems in terms of rate, quantity, or both as a result of human interventions or natural dynamics such as disturbance can have important consequences for ecosystem sustainability and long-term population persistence. Quantitative characterization of spatial patterns and their rates of change in natural environments is essential to understanding ecological processes and to inform sustainable management techniques that aim to minimize degradation and alteration of ecosystem dynamics [3].

Spatial pattern, or simply spatial structure, refers to a quantifiable attribute of a spatial context. General definitions of the word pattern include a simple definition such as a distinctive or regular “form” or “order,” or a feature that is repeated with some degree of “regularity” [4]. Recently, Wagner and Fortin [5] defined the more general term “spatial heterogeneity” as spatially structured variability in a property of interest. Both exogenous environmental (e.g., edaphic variability, elevation, climate) and endogenous ecological (e.g., species interactions, pollen, and seed dispersal) processes generate spatial structure. Each of these two types of spatial processes can produce spatial pattern in multiple forms and scales (Fig. 1). The simplest form of spatial pattern is a simple gradient (Fig. 1a). Spatial



Ecosystems and Spatial Patterns. Figure 1

Spatial heterogeneity as a series of additive processes resulting in additive spatial patterns to which individual organisms (here represented points) may respond. Some or all of the different types of spatial heterogeneity may be present in any given landscape

structure can also be present in the form of patches, linear features, and points and can be superimposed on a gradient (Fig. 1b–f). When biological spatial structure is mostly responding to environmental conditions such as those depicted in Fig. 1, the resulting spatial structure is said to have spatial dependency to the environmental factors. However, when spatial structure emerges as a result of interactions among ecological processes, the pattern is said to be spatially autocorrelated [5].

Spatial patterns and heterogeneity can also be defined using spatial and topological characteristics. These characteristics can include, but are not limited to, a pattern's intensity, autocorrelation, degree of clustering, variability, and scale, which itself includes spatial grain and extent [6]. Importantly, a single pattern summarized using different characteristics can result in different interpretations of the processes behind that pattern [7, 8].

Because spatial pattern analysis is often interested in inferring the processes that created them, it is important to recognize that any single observed pattern represents but one realization of the stochastic process(es) that generated it [7, 9]. By acknowledging that an observed pattern is but a single “snapshot,” its temporal dimension is recognized and that under different circumstances, the patterns seen may not be exactly the same. Hence, a main objective of studying spatial pattern is to try to tease apart stochastic processes and the patterns they create from their spatiotemporal conditionalities.

In addition to being driven by processes that are stochastic, patterns emerge as a result of multiple processes that operate at different spatial and temporal scales [10, 11]. These processes can be biotic or abiotic and are usually interconnected through dynamic, and occasionally nonlinear, feedback loops. For example, emergent spatial pattern following forest fires is conditional on the initial distribution of forest fuels as well as fire-weather conditions [12]. Similarly, patterns in forest vegetation composition are often related to patterns in abiotic factors such as moisture, drainage, and soil conditions. Patterns in the genetic composition in animal populations have also been shown to be influenced by the environmental variation (e.g., suitable vs. unsuitable habitat) between sampled populations [13]. These relationships are often nonlinear as

the patterns that result from the interactions among pattern-generating processes tend to be different than any one process on its own [14, 15].

Novel spatial patterns created by contemporary anthropogenic processes have uncertain consequences for natural ecosystem dynamics. Anthropogenic processes including deforestation, development, land use change, and climate change do not replace natural processes, but have the capacity to interact with and alter them. As such, a significant question in modern ecology and ecosystem science is that of what are the effects of such novel patterns and processes on natural, or historical, system dynamics [16–18]. Not only do new sources of spatial variability influence natural dynamics through changing the patterns to which natural processes respond, but they also can alter the processes themselves. For example, with regard to forest fire dynamics, this is true where forest composition has been changed due to fire suppression and management (pattern change) and fire frequencies are increased due to increased ignitions near roads or changes in local weather patterns (process change). Similarly, with regard to animal population dynamics, movement and dispersal may be impeded through habitat loss and fragmentation (pattern change) and habitat loss can have an absolute effect on effective population size, rates of dispersal, and genetic variability (process change). Sophisticated spatial statistical analyses are required to begin to disentangle the contributions of different processes to observed spatial patterns to understand how best to manage natural systems to safeguard against further habitat-related losses to biodiversity [19].

Here the causes and consequences of spatial patterns in terrestrial forest ecosystems are reviewed with particular emphasis on patterns of forest vegetation generated through landscape level disturbance processes. Spatial patterns in forest vegetation are both ecologically and economically important in that they are directly relevant to wildlife habitat supply, timber supply, future disturbance dynamics, and represent future challenges to forest and land managers. Uncertainty regarding future disturbance dynamics, in particular fire and insects outbreaks, in the context of global climate change makes investigations into disturbance interactions and potential long-term consequences for ecosystem spatial structure and functioning particularly relevant.

Forest Ecosystems

In North American forests, disturbance processes generally include landscape level fires, insect outbreaks, forest management (i.e., logging), and fine-scale local disturbances such as windthrow and fungal diseases. The patterns created through the interactions among disturbances can have important economic and ecological consequences. For example, Stadler et al. [20] demonstrated that hemlock woolly adelgid (*Adelges tsugae*) infestations in New England can affect both fast and slow ecosystem dynamics, nutrient cycling dynamics in the short term, and landscape-scale patterns of forest composition in the long term. Similarly, compounded disturbances (e.g., fire and logging) in the eastern boreal forest can result in alternate forest states [21], which can have consequences for biodiversity conservation. Economically, it has been clearly demonstrated that forests under risk of disturbance, either through fire or insect outbreaks, required longer rotation periods to accommodate for the losses [22].

Logging, fire, and insect outbreaks represent disturbance processes that revert forest stands to early seral stages. Succession describes processes of forest recovery, regeneration, and change that vary in response to different disturbances. Although multiple processes generate forest spatial heterogeneity, not all influence it in the same way. Spatial disturbance legacies vary in terms of shape, size, intensity, boundary characteristics, influence on forest succession, and effects on forest age structure [18, 23–25]. The interactions among processes, or more properly, interactions among current disturbance and existing spatial legacies, create and maintain heterogeneous forest landscapes. This cascade of effects and constraints creates mutual dynamic feedbacks among patterns (spatial legacies) and spatial processes (disturbances) [26, 27] with important consequences for ecosystem dynamics.

Different forest disturbances create different forms of spatial structure. Indeed, each disturbance imposes its own unique “spatial signature” on the landscape that also has different temporal characteristics contingent on a disturbance’s interaction with succession (Fig. 2). Fires, for example, tend to produce relatively discrete patches that occur over a short time frame and vary in terms of the residual forest structure that is left behind [28]. Logging is somewhat similar to fires in

that the patches created are discrete and occur over short time frames and forest managers have control over the scale and amount of residual structure. Insect disturbances, such as outbreaks of spruce budworm (*Choristoneura fumiferana*), forest tent caterpillar (*Malacasoma disstri*), and the mountain pine beetle (*Dendroctonus ponderosae*), are less discrete and tend to produce more complicated spatial structure and continue to affect forest structure at a given location for multiple consecutive years [29].

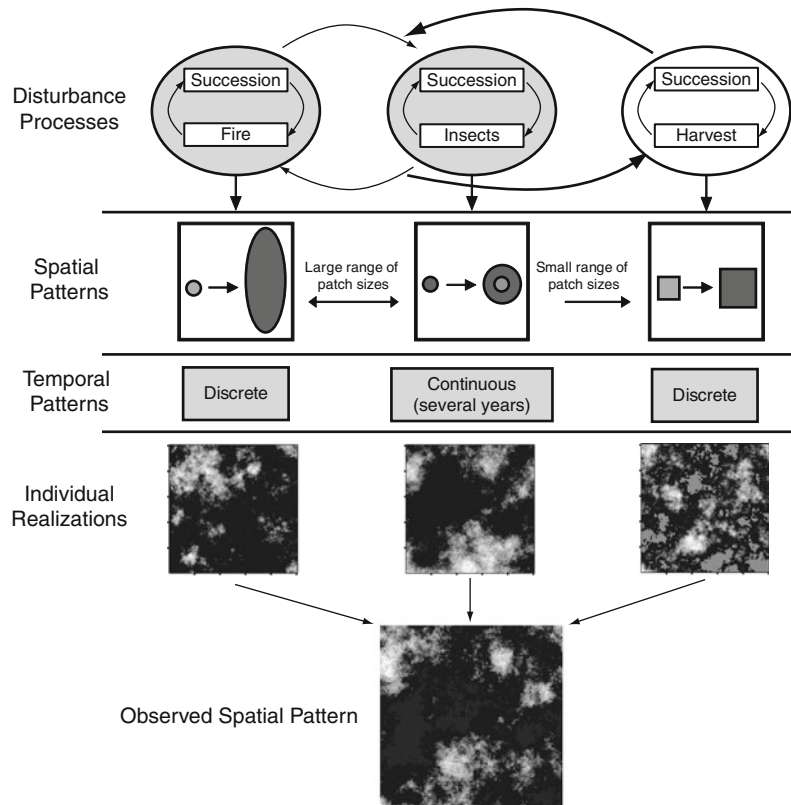
Each disturbance also has a unique relationship with forest regeneration processes [30] such that forest succession is tightly coupled to the type of disturbance that reinitiates stand development. These relationships determine future forest structure. Historically, fire and insects were the main disturbances in North American forest systems. Adaptations to disturbance such as serotiny in pine species (e.g., jack pine; *Pinus banksiana*) and advanced regeneration in the understory of spruce (e.g., *Picea spp.*) and balsam fir (*Abies balsamea*) stands that maintain spruce budworm host availability over time [31] are evidence of this dynamic feedback between disturbance and succession. The spatial patterns created through forest management and their influences on forest succession in turn influence future forest disturbances dynamics [32]. Spatial pattern analysis is important to better understand the effects of human activities on natural disturbance dynamics.

Sources of Heterogeneity

Understanding the nature and consequences of spatial heterogeneity in ecosystems requires an understanding of the processes that generate this heterogeneity. In this section, different types of spatial heterogeneity and how different types of processes may give rise to complex spatial patterns are described. The consequences and potential challenges involved in identifying the relative contributions of these different and frequently interacting processes are discussed next [6].

Levels of Organization

Processes that generate spatial heterogeneity can be classified into a hierarchy of spatial processes that operate at different levels of ecological organization: (1) individual, (2) population, (3) community, and



Ecosystems and Spatial Patterns. Figure 2

Spatial and temporal scales of forest disturbances. Interactions among disturbances are dependent on the unique successional responses to each disturbance (row 1). Historical forest systems were governed by interactions mainly between fire and insects (arrows) although presently, logging also interacts with these historical processes. Columns show the unique spatial and temporal attributes of each of the three main boreal forest disturbance agents: fire, insects (i.e., SBW), and harvesting (i.e., logging). These different spatial features result in different realization, or spatial signatures, of each disturbance (row 4). Interactions among these different processes produce a single observed spatial realization of spatial structure that contains elements of each of the different processes (row 5). Observed patterns contain elements of all three main processes; the objective of spatial analyses is to begin to tease apart the relative contributions of different processes to observed spatial pattern (row 6)

(4) landscape/ecosystem. Individual processes include organism dispersal and habitat selection; population processes can include demographic dynamics as well as immigration/emigration; community level processes are highly relevant to natural disturbance dynamics and can include successional changes and rates of species turnover. Examples of landscape/ecosystem level processes include disturbance, climate change, and migration.

Processes within this organizational hierarchy are not necessarily independent and can influence each other among levels. Such interactions can influence emergent patterns due to potential cross-scale

interactions and amplifications [33] and can also further complicate efforts to identify clear cause-and-effect relationships. A recent example of such cross-scale amplification in a forest ecosystem can be found within the mountain pine beetle (*Dendroctonus ponderosae*) system of Western Canada, where the recent outbreaks of the lodgepole pine infesting beetle have affected an unprecedented millions of hectares [34]. Here, the local dynamics of population control by host tree defenses were overcome when population numbers increased dramatically due to persistent warmer temperatures in the early 2000s.

The interactions between these local- and landscape-level processes are thought to have led to a positive feedback that allowed the outbreak to expand as much as it did [34].

Interactions Among Spatial Processes

Spatial processes within ecosystems interact with each other directly as well as with the spatial legacies of previous, historical processes. In this way, spatial patterns and processes are connected through a dynamic and persistent feedback loop [32]. Spatial legacies can be thought of as a form of spatiotemporal connectivity among disjunct spatial processes or events that are mediated by forest succession and aging. The legacies of historical processes as represented by contemporary patterns can have long-lasting and significant impacts on biodiversity [16, 35] and efforts to sustainably manage forest ecosystems [23, 36].

Spatial legacies can be defined at different scales and describe persisting features within a stand, landscape, or ecosystem. The term “legacy” can refer to fine-scale structural complexity following windthrow [37], landscape level forest age structure [18, 38], disturbance-mediated seed availability [39], and residual forest structure following fire [39, 40]. Because patterns of historical land use can influence contemporary ecosystem composition, configuration, and ecosystem process dynamics long after the actual event [41], a better understanding of spatial legacies and their influence on ecosystem dynamics and landscape change over time is needed and requires novel spatial and temporal methods of investigation and analysis.

From the perspective of sustainability, spatial disturbance legacies, including those created through human activities, represent future ecosystem patterns and future challenges for sustainable management. Gustafson et al. [42] showed that new forest harvest goals are not easy to achieve due to existing conditions when examining shifting forest management rules. Wallin et al. [38] demonstrated that shifts from a dispersed to an aggregated harvest pattern did not immediately result in a change in forest attributes such as patch size and edge density. Instead, new harvest rules had to work around the legacies of previous patterns, and original patterns were enforced. Similarly, Gustafson and Rasmussen [43] found that when varying parameters in a harvest simulation model, the

persistent legacies of previous harvest patterns resulted in timber harvest shortfalls. Using a simulation approach, James et al. [18] demonstrated that legacies in forest age structure created through forest management can persist for over 100 years. Ecologically, the consequences of these legacies interacting with new disturbances can result in greater system variability and gradual ecosystem degradation [43–45] or alternative stable states [21, 46].

Multiple Spatial Scales

Inferring the characteristics of spatial processes through analysis of spatial pattern is a central goal of most ecological studies as it is often very difficult to analyze the processes of interest directly. This can be particularly challenging when several types of pattern (Fig. 1) and underlying processes are present (Fig. 2). The challenge resides in the fact that the single observed pattern is an amalgamation of these multiple processes interacting with existing spatial structure and historical legacies; the functional relationships that connect these contributors to pattern are largely uncertain (Fig. 2). When analyzing the spatial structure of sampled data, it is not easy to disentangle the key spatial scales, and therefore processes, that act on the data. However, in the last decade, hierarchical decomposition methods (multi-scale ordination [47]; PCNM [48]; wavelets, [49] more detailed below) have been developed to identify the spatial scales at which data are most strongly structured and to decompose the data on the basis of scale-specific variances. Beyond simply describing patterns and the scales at which they are structured, it is also important to have a priori hypotheses about which scales and processes are the most relevant for the questions under study as these methods could reveal many patterns and spatial scales, many of which may not be of relevance [8]. Spatial pattern analysis will be more effective at describing underlying processes when used in an explicit and informed hypothesis testing framework.

Ecological Consequences of Spatial Heterogeneity

The consequences of changes in spatial pattern in forest landscapes are easily confounded by absolute losses in wildlife habitat [19]. That is, although both forest

composition and configuration are important, issues related to configuration are only relevant below a critical threshold of forest amount (usually 20–30% of area; [50]). Above the critical threshold, the landscape generally remains “connected” and organisms or disturbances can spread in the landscape [51]. Below such a critical threshold, species’ response to the amount of habitat area is nonlinear as most species do not have enough habitat to meet their needs. Surrounding habitat quality (composition) and configuration become more important for local population persistence in this case. Moreover, fragmented landscapes with various landcover types can impede species abilities to move from habitat patches to another [52]. For example, nesting birds do not cross forest gaps larger than 25 m [53]. Impediments to movement across landscapes can influence population dynamics [54] as well as genetic heterogeneity [55], both of which affect the probability of population persistence.

Spatial Analyses

There are three main approaches to investigating the different aspects and consequences of spatial heterogeneity. Spatial statistics, landscape metrics, and statistical modeling, all approach the question of identifying spatial pattern in ecosystems in a slightly different way [5]. Owing to the varied history of approaches to studying spatial patterns including methods and concepts drawn from a diverse set of disciplines such as geography, geostatistics, and ecology, only the key concepts related to the most commonly used methods that quantify spatial structure within and among ecosystems are presented. Before describing the different types of analysis, some fundamental issues related to the spatial analysis of data are presented.

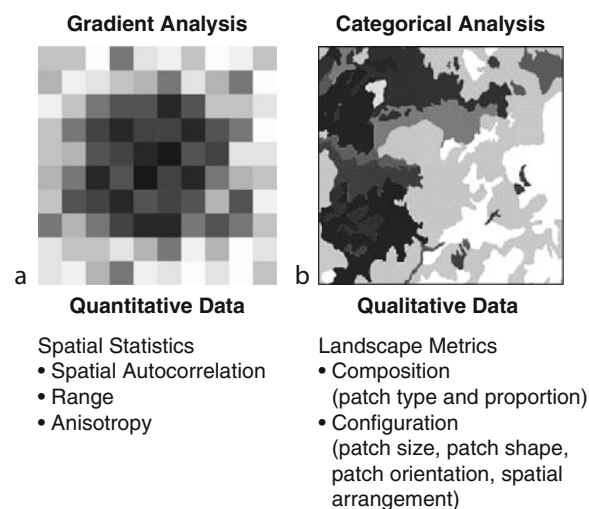
Assumption of Stationarity

To infer spatial pattern from samples spatial statistics require that the area under study is governed by the same underlying process (i.e., the assumption of stationarity; [7]). As it is often impossible to be sure that the underlying process is stationary, one needs either to assume it or to determine whether or not the observed data are stationary (i.e., their statistical properties such as mean, variance, isotropy do not vary with spatial distance). Nonstationary processes may arise

when more than one process is present and that these multiple processes may be acting at different spatial or temporal scales. Yet, in most forest ecosystems, processes interact with one another, resulting in unique types and scales of spatial pattern which violate the assumption of stationarity. In such circumstances, it is required to first identify stationary subregions within such a larger spatial context. A few spatial analysis methods do not require the stationarity such as lacunarity analysis, local quadrat variance methods, and wavelets [7]. It is worth noting that these types of analyses although different and originating from different developmental histories are quite similar to one another mathematically [56].

Data Type

Spatial pattern within ecosystems can be represented using categorical or continuous data depending on the nature of the variable under investigation. Each type of data requires different methods of analysis (Fig. 3; [7]).



Ecosystems and Spatial Patterns. Figure 3

Spatial analysis can be undertaken on different types of spatial data. **(a)** Raster-based quantitative spatial data (e.g., forest height, basal area, NDVI) that can be analyzed using spatial statistics to determine the intensity, spatial range, and directionality (anisotropy) of the spatial pattern. **(b)** Categorical and qualitative forest data (e.g., species, stand age) require a different analytical approach that typically includes landscape pattern metrics

Categorical data can be described by the amount and configuration of the different discrete types on the landscape. Examples of categorical spatial data include forest type and age, or classified habitat patches. Both amount and configuration can be described in numerous ways using landscape pattern metrics [9, 57, 58]. Continuous data requires more subtlety in describing patterns and can include variables such as soil moisture, forest basal area, or remotely sensed reflectance indices (e.g., NDVI). Composition of continuous variables can be described using the density distribution of the variable and configuration is usually described by a spatial covariance function that captures the strength, directionality, and scale of autocorrelation of the variable [5, 7, 59].

In addition to the categorical/quantitative dichotomy of data types, patterns can be described using different geometric topologies of spatial features or units: the vector format (points, lines, polygons) and the raster (i.e., pixel) format [60]. Representation of spatial structure in one form of data does not preclude use of another form. For example, annual polygons of insect defoliation data can be converted into raster form and analyses can be undertaken using time series of raster values at a specific location [61]. Furthermore, binary rasters (presence/absence) can be converted to continuous rasters by increasing the cell size and counting the number of “presence” pixels surrounding a focal pixel using 4-neighbour, 8-neighbour, 16-neighbour rules and assigning this value to the new, larger pixel.

Raster data can be used to represent any continuous variable. In contrast to point data, in raster data types, the information fully covers the extent of the study area. There is also a unique grain (or cell size) to each raster “pixel” that determines the subarea of continuous space that is discretized by the raster cell. The selection of raster grain can have important consequences to the results of spatial analyses [6]. Raster data can be used to represent any number of spatial variables relevant to disturbance ecology including, but not limited to, tree species [62], stand age [18, 63], basal area [64], insect damage [61], and number of fire occurrences [65]. The continuous coverage of raster data makes it amenable to many different analytical techniques such as local quadrat variance, lacunarity, and wavelets [56].

Spatial Analyses within an Ecosystem

Ecological variables that are geographically distributed in space and time tend to be more similar when compared close together [66]. Autocorrelation is a feature of most data and can be quantified by the degree of self-similarity or dissimilarity in a variable between pairs of locations at a given distance apart (i.e., spatial lag determined in terms of equidistant classes). Note that spatial statistics on their own cannot differentiate between spatial dependence to environmental factors and spatial autocorrelation due to ecological processes; only prior knowledge and multiple testing can differentiate between these two sources of spatial structure.

Spatial Description of the Pattern

The objective of many spatial statistics is often to characterize to what degree spatial data are autocorrelated, if they are oriented in a particular direction (anisotropy), and at what scale. As these spatial statistics have been thoroughly reviewed elsewhere [7, 58], we focus on three topics (1) methods of spatial pattern analysis devoted to identifying structure in point data, (2) methods of spatial analysis that are devoted to identifying structure and pattern in two-dimensional raster (pixel) or polygon data, and (3) methods of spatial analysis explicitly concerned with identifying the scale, or scales of structure that are present in either point or two-dimensional data. Both the data types discussed can be examined in uni-, bi-, and multivariate contexts and can include either categorically or continuously measured variables.

Point Pattern Analysis

Spatial point processes describe phenomena that produce events represented as points in space [67, 68]. The objective of point pattern analysis is to determine whether the distribution of events (points in space) is more or less spatially aggregated than is expected by chance and tests the null hypothesis of complete spatial randomness. The use of complete spatial randomness assumes that the underlying process is the same over the study area (i.e., stationarity). When it is not the case, the study area is said to “inhomogeneous” such that significance cannot be achieved using a single process such as complete spatial randomness. Modified

statistics and corrections have been developed to account for inhomogeneity within the study area [69, 70]. Point pattern analysis also assumes complete census of all point occurrences in the study area [7]. Fortin et al. [71] showed that the significance of spatial aggregation estimations is biased only when a subset of sample points is used rather than the entire set of points in the study area.

The most commonly used method of point pattern analysis is Ripley's K statistic [72]. New statistics have also been developed to compute local estimates [68]. Ripley's K statistic, and derived statistics, can be applied in one-, two-, or three-dimensional space to compare the degree of aggregation of points [68]. In cases where two (or more) point processes are operating, it may be of interest to assess whether one process influences the other and whether points of different types tend to cluster together. Examples of ecologically relevant spatial point processes include fire occurrence, plant occurrences [67], or the distribution of animal nesting or denning sites [72]. The bivariate, or cross- K , Ripley's K test assesses whether the co-occurrence of two types of points is clustered together more or less than is expected by chance [58]. Using this technique, Lynch and Moorcroft [73] examined co-occurrence of fire and insect outbreaks and found that contrary to expectation, insect-caused forest mortality does not increase the risk of forest fire.

Spatial Autocorrelation

Often, a researcher is interested in determining the scale and strength of spatial autocorrelation of a variable as well as whether there is a directional trend in the data (i.e., anisotropy). This can be achieved using spatial autocorrelation coefficients such as Moran's I which computes the product of the deviations of the values of the variable to its average according to various distance intervals (lags, classes) standardized by the variance at that spatial lag [7]. Moran's I behaves like a Pearson's correlation coefficient such that the null hypothesis is the absence of spatial autocorrelation, positive autocorrelation (mostly a short distance) indicates that values have comparable values, while negative values indicate that the values are very dissimilar. Moran's I assumes that the underlying process is the same over the entire study

area (i.e., stationarity). Hence, the spatial autocorrelation coefficients computed at various distance classes are average values. Spatial autocorrelation in this sense can be referred to as a global spatial statistic that describes an attribute of the data over the entire study area [7]. Significance of each coefficient can be computed based on an asymptotic t or randomization procedure. In either case, stationarity is required. Moran's I is very sensitive to skewed data as the mean will be biased and in consequence all the deviations values based on it will also be biased. It is therefore recommended to check the distribution of the data before computing spatial autocorrelation and if needed transform the data to obtain a symmetric distribution.

Measures of spatial autocorrelation are also sensitive to sample size. When autocorrelation is estimated using too few locations, (e.g., <30 positions), spatial patterns may not be detected, even though present. Similarly, depending on the spacing among sampling locations, there will be different numbers of paired comparisons at each lag distance. Typically, there are few pairs at short distances due to the edge effects of the edge of the study area (no locations outside to compare too), most of the pairs at intermediate distances and very fewer at large distances (because of the overall size of the study area). To mitigate these unequal numbers of pair per distance lag, it is recommended to focus on lags distances equal to the length of the first half (up to two thirds) of the smallest edge of the study area as most spatial autocorrelation occurs at short distances and that the probability of detecting it is also highest in the first spatial lag [7].

A plot of spatial autocorrelation coefficients against distance lags is called a spatial correlogram. If the lags are based on distance only, the correlogram is said to be an omnidirectional correlogram. When the data contain directionality, the omnidirectional correlogram cannot reveal it and may, in fact, "mask" it. To detect the presence of anisotropic spatial pattern (i.e., not having the same sill and range according to direction), the samples need to be divided by distance class as well as direction angle range (usually 0° , 45° , 90° , 135°) to produce a set of directional correlograms.

In areas where several processes influence ecological data, Moran's I that assumes stationarity cannot be used. Instead, local indicator of spatial aggregation statistics, LISA (e.g., local Moran, local Getis), can be

used as they are computed at each sampling location and allow the identification of subareas that have similar high (“hot spots”) or low (“cold spots”) values [7].

Geostatistics

Spatial structure can be determined in terms of spatial autocorrelation as presented above or as spatial variance according to distance as computed using variograms which are part of the family of spatial statistical methods known as geostatistics [7, 58]. Variograms represent a global method of scale-specific analysis that has been used extensively in ecology to analyze spatial patterns [58]. Variograms model the relationship between lag distance and semivariance and can be calculated using continuous raster or point data. Semivariance is calculated as the sum of the squared differences between pairs of locations separated by a given lag distance divided by twice the number of pairs of locations at that particular lag distance [7]. From the observed, or empirical, variogram, three parameters can be estimated to fit a theoretical variogram: (1) range, or scale at which distance does not affect the estimate of variance, (2) sill, or the variance of the data, and (3) nugget effect, which represents the variability in the data that is not accounted for by spatial structure [58]. Theoretical variogram models can identify whether there is a directional trend in the data, that is, anisotropy as the spatial autocorrelation values. In addition to describing the attributes of spatial structure, geostatistical models can be used to “krige” (i.e., spatially interpolate) data [58] and to simulate spatial patterns using a chosen variogram model, process model, and parameter estimates [74].

Spatial Scale and Scaling

In addition to understanding the type (e.g., trend, patch) and strength (e.g., degree and distance) of a spatial pattern, it is useful to identify the spatial scale at which such patterns are present. Because patterns are the result of multiple processes that each have their own unique scales of spatial structure [75], disentangling the relative contributions of these processes and assigning relative importance to them and the scales at which they operate are of fundamental importance to ecology and improving the understanding of complex systems

[11, 16]. Furthermore, the identification of the relative contributions of different processes and scales to observed patterns is necessary for understanding cross-scale interactions [33] which is necessary to make reliable predictions of system dynamics, ostensibly the objective of any spatial analysis [76].

As stated above, spatial pattern describes a “quantifiable attribute of a spatial context.” Scale-specific analysis identifies the spatial scale at which that attribute is structured establishes its specific context. With regard to forest disturbance dynamics, for example, individual forest stands may seem unstable through the processes of destruction and renewal through disturbances such as fire, but the larger forest landscape (i.e., collection of stands) is in fact stable with respect to the proportion and relative configuration of the different stand types. This is what is meant when disturbance-mediated forest systems are described as a shifting mosaic [77]. Different conclusions would be drawn about forest stability and resilience (*sensu* [78]) depending on the spatial or temporal scale of investigation.

Scale generally describes the spatial extent, grain, and thematic resolution of a set of data [6]. However, scale can also be used to refer to a level within an organizational hierarchy to which such data pertain, such as a population, or community, or ecosystem [79]. It is important to note that scale in this latter sense is not directly equivalent to the former; scaling up, that is, increasing from local to a broader extent, or aggregating data from a fine to a coarse scale may move the analysis into another level or an organizational hierarchy, but not necessarily [80]. Although scale is best thought of as a feature of the phenomena of interest, it can also be a feature of sampling scheme imposed by a researcher, or the methods of analysis applied [6]. All of these features can influence a researcher’s ability to identify scales of structure in spatial data and to make meaningful inferences regarding the underlying processes. It is therefore important to distinguish structure that is emergent from the data from those related to sampling or analytic scales (i.e., arbitrary scales; [10]), as such *a priori* scales may have little to do with the actual scale of structure in the ecological phenomena of interest [76].

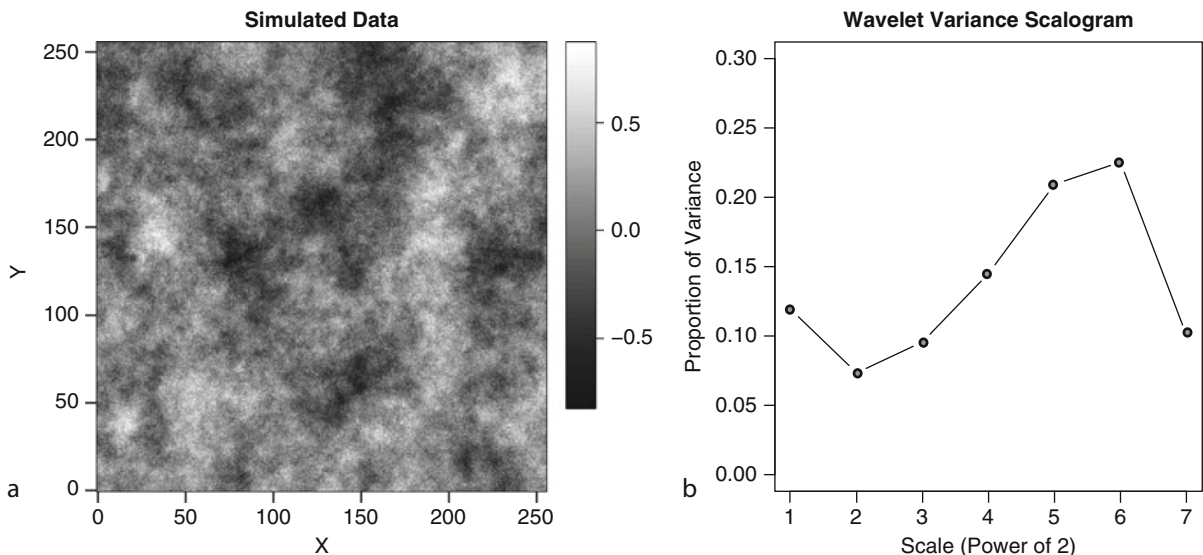
The ability to identify meaningful scales of spatial structure depends on the methods used and the type of data being analyzed [77, 81]. Methods differ in their

ability to identify local vs. global scales of pattern. Global methods of scale-specific analysis summarize spatial pattern at a single scale and generally assume that the underlying processes are stationary. Examples of global methods of analysis include variography [58], spectral (i.e., Fourier) analysis [82], and global measures of spatial autocorrelation such as Moran's I and Geary's c [7]. Multi-scale methods of analysis identify both global and local scales of structure, can assign relative importance to different scales, and do not assume stationarity. Instead, such methods can be used to identify boundaries and scale-specific stationary subregions within a larger spatial context [7, 74]. Multi-scale methods of analysis include lacunarity analysis [83], wavelets [74, 81, 84], distance-based Eigenvector methods (e.g., PCNM; [85]), and local spatial statistics [86].

Wavelet analysis is a particularly powerful method of local spatial analysis that can be used to decompose continuous data into its scale-specific components [87, 88]. A proportion of the total variance in the data set is associated with each level of the decomposition through use of the wavelet variance [84] and the

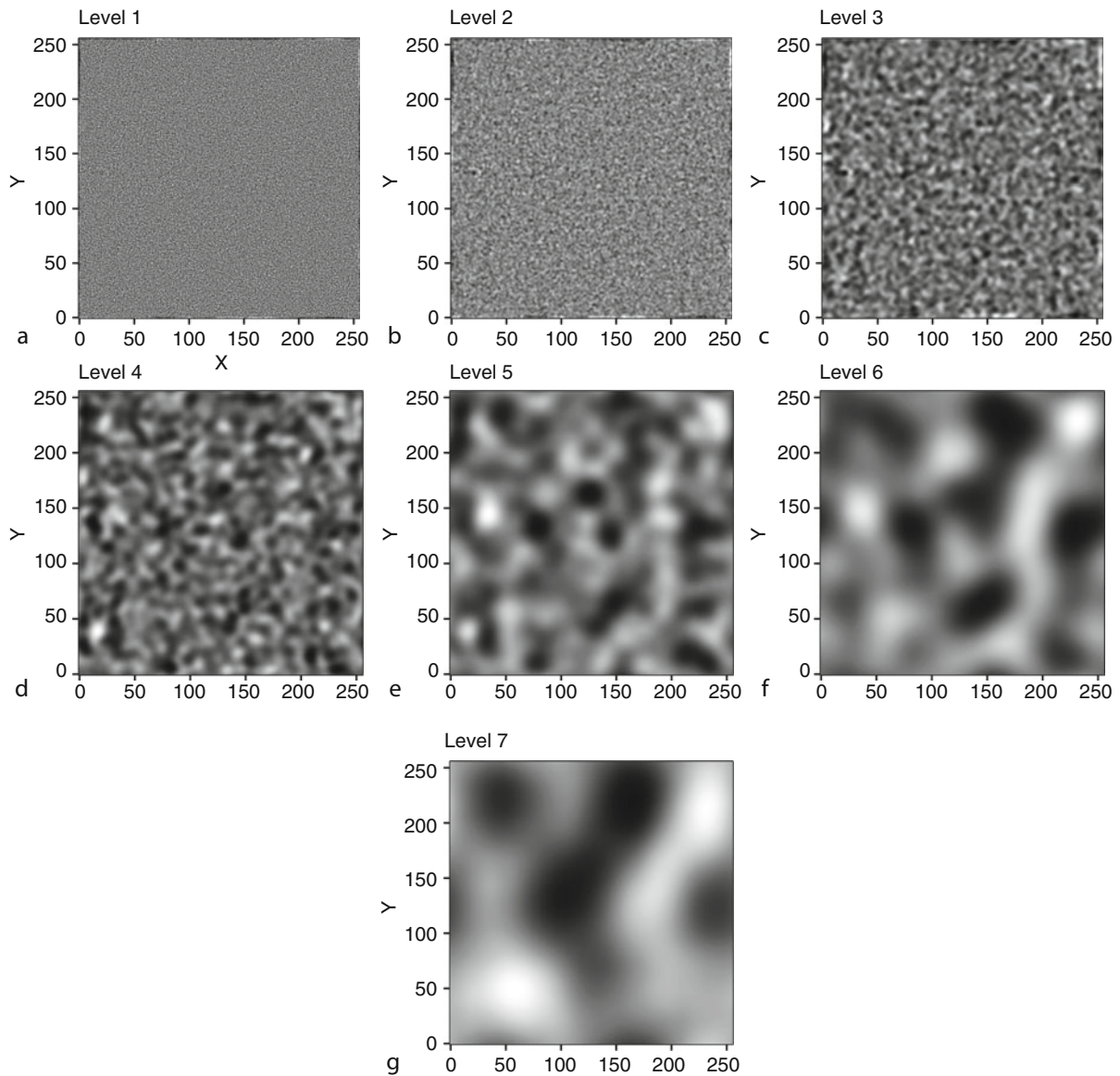
relative contributions of different scales to overall structure can be assessed and visualized as a scalogram (Fig. 4b). Each of these scales of pattern can then be isolated using a multi-resolution decomposition (Fig. 5; [88]). Under conditions where observed spatial pattern is assumed to be the result of multiple interacting processes, such data decomposition provides an opportunity to assess the relationship among processes and individual scales of spatial structure present in the data. In combination with the scalogram (Fig. 4b), the relative importance of these different scales can also be determined and further analyses can be restricted to only those spatial layers that correspond to the scales of interest. Isolated scales of spatial pattern can then be examined independently or used as scale-specific predictors in further statistical analyses [89].

Whereas Fourier analysis assumes that observed patterns can be described as a sum of sine waves of different frequencies, wavelet analysis identifies global and local structure at different scales using a local wavelet template that can take on a wide variety of shapes and forms [90]. The majority of wavelet



Ecosystems and Spatial Patterns. Figure 4

Example of a hierarchical multi-scale decomposition of two-dimensional, quantitative data using wavelets. **(a)** Simulated spatial data. (Data were simulated using an exponential variogram model with the following parameters: Sill = 1; Range = 40; Nugget = 0.1). **(b)** Scalogram that summarizes the proportion of total variance in the original data associated with each scale of the decomposition. Wavelet decomposition was accomplished using a maximal-overlap discrete wavelet transform (MODWT; Percival and Walden [109])



Ecosystems and Spatial Patterns. Figure 5

Back-transformed images from the decomposition of the data in Fig. 4a. Each panel shows the spatial structure at an individual isolated scale

applications have been in the analysis of temporal signals to identify periodicity in things such as climatic variability [91] and epidemiological time series [92, 93]. However, spatial applications of wavelet analysis in ecology continue to be developed and have been used to investigate one-dimensional forest canopy gap structure [87], vegetation reflectance [89], two-dimensional structure in grassland productivity [81],

tree crown identification [94], and the significance of spatial structure in forest basal area [74].

When data are not sampled in a continuous way (i.e., they are irregularly spaced), multi-scale decompositions can be performed using spatial eigenfunction analyses such as principal coordinate analysis of neighbor matrices (PCNM) and Moran's eigenvector maps (MEM) [48, 95, 96]. These methods model spatial structure in

a multivariate framework using distance matrices where the spatial coordinates of the sampled sites are converted into a set of synthetic spatial variables that represent spatial structure at different spatial scales. These synthetic variables can then be used in combination with constrained ordination methods (e.g., canonical correspondence analysis and redundancy analysis; [95]) to identify the spatial scales at which data are dominantly structured. Jombert et al. [97] also proposed a multi-scale pattern analysis (MSPA) to determine which of these many scales are the most relevant to use as spatial predictors in subsequent analyses such as partial ordination or multiple regression [85].

Spatial Analyses Among Ecosystems

The ecotonal interfaces between ecosystems are important to delineate as they are the locations where the exchange of nutrients and species turnover occurs [7]. It is also important to determine not only the boundary location between ecosystems, that is, where one system begins and another ends, but also its width (i.e., sharp/line or gradual/zone) [98]. There are two different types of methods that can be used to determine the interface between ecosystems: by creating spatial clusters or by detecting boundaries [99]. In either case, the sampling design used to collect the data is crucial: The determination of a boundary (i.e., area of high rate of changes in the values of variables, hence a heterogeneous area) is relative to the two adjacent spatially homogeneous ecosystems. Therefore, the sampled data should cover enough of both ecosystems such that their interface can be detected.

Spatial Clustering

To delimit boundaries between ecosystems, spatially homogeneous clusters can be determined based on the degree of similarity of sample attributes and their spatial adjacencies [66, 99, 100]. The degree of similarity can be based on commonly used clustering algorithms (agglomerative, *k*-means, fuzzy logic, etc.) and adjacency can be based on network connectivity algorithms (e.g., nearest neighbors, minimum spanning tree, Gabriel network, Delaunay network; [7]). Spatial clustering provides the membership of each location to a spatial cluster and therefore, as a by-product, identifies boundaries among clusters. However, clustering

procedures do not provide any information about the location and width of the identified boundaries between clusters. When information about location and width are needed, other methods should be used such as boundary detection methods ([99]; but see [101]).

Spatial Boundary Analyses

Ecological boundaries can be defined as areas of high rates of change or large absolute differences between adjacent locations [102]. Boundary detection methods [98] include edge detection algorithms (Laplacian, Canny, Sobel, Monomier, etc.), wavelet analysis [74, 81], and wombling (lattice, triangulation, categorical). The two former families of methods require the data to be in a contiguous fashion (grid) without any missing values while the latter one can either be used with contiguous or irregularly sampled data. All methods compute the magnitude of rate of change between adjacent locations over the entire extent of the study area. To determine which rates of changes are significantly higher than the others, different statistics and significance tests having been developed. Wombling typically uses arbitrary percentile thresholds of boundary elements to identify significant boundaries [7]. James et al. [74] proposed a series of restricted randomization procedures to test the significance of wavelet boundaries using variogram-based spatial null models. Oden et al. [103] developed boundary statistics to test the cohesiveness properties of the boundaries. Once cohesive boundaries have been detected and tested, subsequent hypothesis testing can be performed by comparing the spatial overlap and movement of boundaries using spatial overlap statistics [7, 104, 105] or polygon change analysis [106].

Future Directions

The detection and characterization of spatial structure is necessary for developing the understanding of ecosystem function and for ensuring sustainable management and use of landscape resources. Indeed, the current scale and pace of anthropogenic influence on the natural environment is without precedent and the ways in which novel human-created spatial patterns interact with and influence natural processes are uncertain. Even more uncertain are the relationships among different scales of spatial and temporal pattern and what such cross-scale

interactions may mean to ecosystem dynamics [107]. Future directions in the analysis of spatial pattern will require approaches and methods that can begin to tease apart the separate scales, both spatial and temporal, of processes that contribute to spatial patterns both within and among ecosystems. These methods should include multi-scale approaches such as the wavelet-based methods described above, tools to assess statistically significant changes over time, and methods that can identify local spatially significant subregions within larger spatial contexts [74]. Meaningful inference in these regards will only be possible if the spatial pattern analysis is applied within a hypothesis testing framework, where competing notions of how individual processes percolate through the landscape to produce pattern can be tested statistically [8]. Finally, increasing availability of remotely sensed data (e.g., NDVI, LiDAR, Quickbird, LANDSAT, MODIS) will allow people to detect spatial patterns at finer spatial and temporal scales over much larger spatial extents than has been previously possible. These huge amounts of data will also require focused, hypothesis-driven questions, the use of data-mining tools [108], and the further development of spatiotemporal statistics.

Acknowledgments

This work was funded by a Killam Postdoctoral fellowship at the University of Alberta to PMAJ and an NSERC Discovery grant to MJF.

Bibliography

Primary Literature

- Lindenmayer DB, Fischer J (2006) *Habitat fragmentation and landscape change*. Island Press, Washington, DC
- Loreau M, Naeem S, Inchausti P et al (2001) Biodiversity and ecosystem functioning: current knowledge and future challenges. *Science* 294:804–808
- Lindenmayer DB, Franklin JF (2002) *Conserving forest biodiversity. A comprehensive multiscaled approach*. Island Press, London
- Epperson BK (2003) *Geographical genetics*. Princeton University Press, Princeton
- Wagner HH, Fortin M-J (2005) *Spatial analysis of landscapes: concepts and statistics*. *Ecology* 86:1975–1987
- Dungan JL, Perry JN, Dale MRT et al (2002) A balanced view of scale in spatial statistical analysis. *Ecography* 25:626–640
- Fortin M-J, Dale MRT (2005) *Spatial analysis*. Cambridge University Press, Cambridge
- McIntire EJB, Fajardo A (2009) Beyond description: the active and effective way to infer processes from spatial patterns. *Ecology* 90:46–56
- Fortin M-J, Boots B, Csillag F, Rempel TK (2003) On the role of spatial stochastic models in understanding landscape indices in ecology. *Oikos* 102:203–212
- Wiens JA (1989) Spatial scaling in ecology. *Funct Ecol* 3:385–397
- Levin SA (1992) The problem of pattern and scale in ecology. *Ecology* 73:1943–1967
- Flannigan MD, Stocks BJ, Turetsky MR, Wotton BM (2008) Impact of climate change on fire activity and fire management in the circumboreal forest. *Global Change Biol* 14:1–12
- McRae BH, Beier P (2007) Circuit theory predicts gene flow in plant and animal populations. *Proc Natl Acad Sci USA* 104:19885–19890
- Veblen TT, Hadley KS, Nel EM et al (1994) Disturbance regime and disturbance interactions in a Rocky Mountain subalpine forest. *J Ecol* 82:125–135
- Radeloff VC, Mladenoff DJ, Boyce MS (2000) Effects of interacting disturbances on landscape patterns: Budworm defoliation and salvage logging. *Ecol Appl* 10:233–247
- Levin SA (2000) Multiple scales and the maintenance of biodiversity. *Ecosystems* 3:498–506
- Cooke BJ, Nealis VG, Regniere J (2007) Insect defoliators as periodic disturbances in northern forest ecosystems. In: Johnson EA, Miyanishi K (eds) *Plant disturbance ecology: the process and the response*. Elsevier, Amsterdam, pp 1–10
- James PMA, Fortin M-J, Fall A, Kneeshaw D, Messier C (2007) The effects of spatial legacies following shifting management practices and fire on boreal forest age structure. *Ecosystems* 10:1261–1277
- Fahrig L (2003) Effects of habitat fragmentation on biodiversity. *Ann Rev Ecol Syst* 34:487–515
- Stadler B, Müller T, Orwig D, Cobb R (2005) Hemlock woolly adelgid in New England forests: canopy impacts transforming ecosystem processes and landscapes. *Ecosystems* 8:233–247
- Jasinski JPP, Payette S (2005) The creation of alternative stable states in the southern boreal forest, Quebec, Canada. *Ecol Monogr* 75:561–583
- MacLean DA, Erdle TA, MacKinnon WE et al (2001) The spruce budworm decision support system: forest protection planning to sustain long-term wood supply. *Can J Forest Res* 31: 1742–1757
- Franklin JF, Forman RTT (1987) Creating landscape patterns by forest cutting: Ecological consequences and principles. *Landscape Ecol* 1:5–18
- Fall A, Fortin M-J, Kneeshaw DD et al (2004) Consequences of various landscape-scale ecosystem management strategies and fire cycles on age-class structure and harvest in boreal forests. *Can J Forest Res* 34:310–322
- Greene DF, Gauthier S, Noe J, Rousseau M, Bergeron Y (2006) A field experiment to determine the effect of post-fire salvage on seedbeds and tree regeneration. *Front Ecol Environ* 4:69–74
- Watt AS (1947) Pattern and process in the plant community. *J Ecol* 25:1–22

27. Turner MG (1989) Landscape ecology: the effect of pattern on process. *Ann Rev Ecol Evol Syst* 20:171–197
28. Bergeron Y, Leduc A, Harvey BD, Gauthier S (2002) Natural fire regime: a guide for sustainable management of the Canadian boreal forest. *Silva Fenn* 36:81–95
29. Boulanger Y, Arseneault D (2004) Spruce budworm outbreaks in eastern Quebec over the last 450 years. *Can J Forest Res* 34:1035–1043
30. Schroeder D, Perera AH (2002) A comparison of large-scale spatial vegetation patterns following clearcuts and fires in Ontario's boreal forests. *Forest Ecol Manage* 159:217–230
31. Baskerville GL (1975) Spruce budworm: super silviculturist. *Forestry Chron* 51:138–140
32. Turner MG (1993) A revised concept of landscape equilibrium: disturbance and stability on scaled landscapes. *Landscape Ecol* 8:213–227
33. Peters DPC, Pielke RA, Bestelmeyer BT et al (2004) Cross-scale interactions, nonlinearities, and forecasting catastrophic events. *Proc Natl Acad Sci* 101:15130–15135
34. Raffa KF, Aukema BH, Bentz BJ et al (2008) Cross-scale drivers of natural disturbances prone to anthropogenic amplification: the dynamics of bark beetle eruptions. *Bioscience* 58:501–517
35. Peterson GD (2002) Contagious disturbance, ecological memory, and the emergence of landscape pattern. *Ecosystems* 5:329–338
36. Kohm KA, Franklin JF (1997) Creating a forestry for the 21st century. Island Press, Washington, DC
37. Foster DR, Knight DH, Franklin JF (1998) Landscape patterns and legacies resulting from large, infrequent forest disturbances. *Ecosystems* 1:497–510
38. Wallin DO, Swanson FJ, Marks B (1994) Landscape pattern response to changes in pattern generation rules: land-use legacies in forestry. *Ecol Appl* 4:569–580
39. Tang SM, Franklin JF, Montgomery DR (1997) Forest harvest patterns and landscape disturbance processes. *Landscape Ecol* 12:349–363
40. McIntire EJB (2004) Understanding natural disturbance boundary formation using spatial data and path analysis. *Ecology* 85:1993–1943
41. Dupouey JL, Dambrine E, Lafitte JD, Moares C (2002) Irreversible impact of past land use on forest soils and biodiversity. *Ecology* 83:2978–2984
42. Gustafson EJ (1996) Expanding the scale of forest management: allocating timber harvests in time and space. *Forest Ecol Manage* 87:27–39
43. Gustafson EJ, Rasmussen LV (2002) Assessing the spatial implications of interactions among strategic forest management options using a Windows-based harvest simulator. *Comput Electron Agric* 33:179–196
44. Paine RT, Tegner MJ, Johnson EA (1998) Compounded perturbations yield ecological surprises. *Ecosystems* 1:535–545
45. Folke C, Carpenter S, Walker B et al (2004) Regime shifts, resilience, and biodiversity in ecosystem management. *Ann Rev Ecol Evol Syst* 35:557–581
46. Schröder A, Persson L, De Roos AM (2005) Direct experimental evidence for alternative stable states: a review. *Oikos* 110:3–19
47. Wagner H (2004) Direct multi-scale ordination with canonical correspondence analysis. *Ecology* 85:342–351
48. Borcard D, Legendre P, Avois-Jacquet C, Tuomisto H (2004) Dissecting the spatial structure of ecological data at multiple scales. *Ecology* 85:1826–1832
49. Keitt TH, Urban DL (2005) Scale-specific inference using wavelets. *Ecology* 86:2497–2504
50. Andrén H (1994) Effects of habitat fragmentation on birds and mammals in landscapes with different proportions of suitable habitat – a review. *Oikos* 71:355–366
51. Gardner RH, Milne BT, Turney MG, O'Neill RV (1987) Neutral models for the analysis of broad-scale landscape pattern. *Landscape Ecol* 1:19–28
52. O'Brien D, Manseau M, Fall A, Fortin M-J (2006) Testing the importance of spatial configuration of winter habitat for woodland caribou: An application of graph theory. *Biol Conserv* 130:70–83
53. Bélisle M, Desrochers A, Fortin M-J (2001) Influence of forest cover on the movements of forest birds: a homing experiment. *Ecology* 82:1893–1904
54. Fahrig L, Merriam G (1985) Habitat patch connectivity and population survival. *Ecology* 66:1762–1768
55. Bohonak AJ (1999) Dispersal, gene flow, and population structure. *Q Rev Biol* 74:21–45
56. Dale MRT, Dixon P, Fortin M-J et al (2002) Conceptual and mathematical relationships among methods for spatial analysis. *Ecography* 25:558–577
57. Gustafson EJ (1998) Quantifying landscape spatial pattern: what is state of the art? *Ecosystems* 1:143–156
58. Cushman SA, McGarigal K (2008) Landscape metrics, scales of resolution. In: von Gadow K, Pukkala T (eds) *Designing green landscapes*. Springer, New York, pp 33–52
59. Cressie NA (1993) *Statistics for spatial data*. Wiley, New York
60. Burrough PA, McDonnell RA (1998) *Principles of geographical information systems*. Oxford University Press, Oxford
61. Candau J-N, Fleming R (2005) Landscape-scale spatial distribution of spruce budworm defoliation in relation to bioclimatic conditions. *Can J Forest Res* 35:2218–2232
62. Wolter PT, White MA (2002) Recent forest cover type transitions and landscape structural changes in northeast Minnesota, USA. *Landscape Ecol* 17:133–155
63. Didion M, Fortin M-J, Fall A (2007) Forest age structure as indicator of boreal forest sustainability under alternative management and fire regimes: a landscape level sensitivity analysis. *Ecol Model* 200:45–58
64. Townsend PA, Foster JR, Chastain RA, Currie WS (2003) Application of imaging spectroscopy to mapping canopy nitrogen in the forests of the central Appalachian mountains using Hyperion and AVIRIS. *IEEE Trans Geosci Remote Sens* 41:1347–1354
65. Wotton BM, Martell DL (2005) A lightning fire occurrence model for Ontario. *Can J Forest Res* 35:1389–1401

66. Legendre P, Fortin M-J (1989) Spatial pattern and ecological analysis. *Vegetatio* 80:107–138
67. Dale MRT (1999) *Spatial pattern analysis in plant ecology*. Cambridge University Press, Cambridge
68. Illian J, Penttinen A, Stoyan H, Stoyan D (2008) *Statistical analysis and modelling of spatial point patterns*. Wiley-Interscience, Chichester
69. Baddeley AJ, Møller J, Waagepetersen R (2000) Non- and semi-parametric estimation of interaction in inhomogeneous point patterns. *Statistica Neerlandica* 54:329–350
70. Wiegand T, Moloney KA (2004) Rings, circles and null-models for point pattern analysis in ecology. *Oikos* 104:209–229
71. Fortin M-J, Dale MRT, Bertazzon S (2010) Spatial analysis of wildlife distribution and disease spread. In: Huettmann F, Cushman SA (eds) *Spatial complexity, informatics, and wildlife conservation*. Springer, New York, pp 255–273
72. Ripley BD (1976) The second order analysis of stationary point processes. *J Appl Probab* 13:255–266
73. Lynch HJ, Moorcroft PR (2008) A spatiotemporal Ripley's K-function to analyze interactions between spruce budworm and fire in British Columbia, Canada. *Can J Forest Res* 38:3112–3119
74. James PMA, Fleming RA, Fortin M-J (2010) Identifying significant scale-specific spatial boundaries using wavelets and null models: spruce budworm defoliation in Ontario, Canada as a case study. *Landscape Ecol* 25:873–887
75. Holling CS (1992) Cross-scale morphology, geometry, and dynamics of ecosystems. *Ecol Monogr* 62:447–502
76. Gardner RH (1998) Pattern, Process, and the Analysis of Spatial Scales. In: Peterson DL, Parker VT (eds) *Ecological scale*. Columbia University Press, New York, pp 17–34
77. Saunders SC, Chen J, Drummer TD et al (2002) The patch mosaic and ecological decomposition across spatial scales in a managed landscape of northern Wisconsin, USA. *Basic Appl Ecol* 3:49–64
78. Holling CS (1973) Resilience and stability of ecological systems. *Ann Rev Ecol Evol Syst* 4:1–23
79. Allen TFH, Hoekstra TW (1992) *Toward a unified ecology*. Columbia University Press, New York
80. O'Neill RV, Krummel JR, Gardner RH et al (1988) Indices of landscape pattern. *Landscape Ecol* 1:153–162
81. Csillag F, Kabos S (2002) Wavelets, boundaries, and the spatial analysis of landscape pattern. *Ecoscience* 9:177–190
82. Keitt TH (2000) Spectral representation of neutral landscapes. *Landscape Ecol* 15:479–494
83. Saunders SC, Broszofsky KD, Chen J, Drummer TD, Gustafson EJ (2005) Identifying scales of pattern in ecological data: a comparison of lacunarity, spectral and wavelet analyses. *Ecol Complex* 2:87–105
84. Dale MRT, Mah M (1998) The use of wavelets for spatial pattern analysis in ecology. *J Veg Sci* 9:805–814
85. Borcard D, Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol Model* 153:51–68
86. Anselin L (1995) Local indicators of spatial association - LISA. *Geogr Anal* 27:93–115
87. Bradshaw RHW, Spies T (1992) Characterizing canopy gap structure in forests using wavelet analysis. *J Ecol* 80:205–215
88. Mallat S (1999) *A wavelet tour of signal processing*. Academic, New York
89. Keitt TH, Urban DL (2005) Scale-specific inference using wavelets. *Ecology* 86:2497–2504
90. Daubechies I (1992) *Ten lectures on wavelets*. SIAM, Philadelphia
91. Torrence C, Compo GP (1998) A practical guide to wavelet analysis. *Bull Am Meteor Soc* 79:61–78
92. Grenfell BT, Björnstad ON, Kappey J (2001) Travelling waves and spatial hierarchies in measles epidemics. *Nature* 414:716–723
93. Cazelles B, Chavez M, De Magny GC, Guegan JF, Hales S (2007) Time-dependent spectral analysis of epidemiological time-series with wavelets. *J R Soc Interface* 4(15):625–636
94. Falkowski MJ, Smith AMS, Hudak AT et al (2006) Automated estimation of individual conifer tree height and crown diameter via two-dimensional spatial wavelet analysis of lidar data. *Can J Remote Sensing* 32:153–161
95. Dray S, Legendre P, Peres-Neto PR (2006) Spatial modeling: a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). *Ecol Model* 196:483–493
96. Griffith DA, Peres-Neto PR (2006) Spatial modelling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology* 87:2603–2613
97. Jombart T, Dray S, Dufour AB (2009) Finding essential scales of spatial variation in ecological data: a multivariate approach. *Ecography* 32:161–168
98. Jacquez GM, Maruca S, Fortin M-J (2000) From fields to objects: a review of geographic boundary analysis. *J Geograph Syst* 2:221–241
99. Fortin M-J, Drapeau P (1995) Delineation of ecological boundaries: comparison of approaches and significance tests. *Oikos* 72:323–332
100. Fortin M-J, Keitt TH, Maurer B et al (2005) Species ranges and distributional limits: pattern analysis and statistical issues. *Oikos* 108:7–17
101. McIntire EJB, Fortin M-J (2006) Structure and function of wildfire and mountain pine beetle forest boundaries. *Ecography* 29:309–318
102. Fortin M-J (1994) Edge detection algorithms for two-dimensional ecological data. *Ecology* 75:956–965
103. Oden NL, Sokal RR, Fortin M-J, Goebel H (1993) Categorical wombling: detecting regions of significant change in spatially located categorical variables. *Geogr Anal* 25:315–336
104. Fortin MJ, Drapeau P, Jacquez GM (1996) Quantification of the spatial co-occurrences of ecological boundaries. *Oikos* 77:51–60
105. St-Louis V, Fortin M-J, Desrochers A (2004) Association between microhabitat and territory boundaries of two forest songbirds. *Landscape Ecol* 19:591–601

106. Sadahiro Y, Umemura M (2002) A computational approach for the analysis of changes in polygon distributions. *J Geograph Syst* 3:137–154
107. Peters DPC, Bestelmeyer BT, Turner MG (2007) Cross-scale interactions and changing pattern-process relationships: consequences for system dynamics. *Ecosystems* 10:790–796
108. Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning: data mining, inference, and prediction*, 2nd edn. Springer, New York
109. Percival and Walden (2000) *Wavelet methods for time series analysis*. Cambridge, Cambridge University Press

Ecosystems, Adaptive Management

CRAIG R. ALLEN¹, JOSEPH J. FONTAINE¹,
AHJOND S. GARMESTANI²

¹U.S. Geological Survey, Nebraska Cooperative Fish and Wildlife Research Unit, School of Natural Resources, University of Nebraska, Lincoln, NE, USA

²U.S. Environmental Protection Agency, National Risk Management Research Laboratory, Cincinnati, OH, USA

Article Outline

Glossary
 Definition of Adaptive Management
 Introduction
 Adaptive Management Today
 Structured Decision Making
 Participatory Active Adaptive Management
 Adaptive Governance
 Adaptive Management and Law
 Conclusions
 Future Directions
 Acknowledgments
 Bibliography

Glossary

Adaptive governance Institutional and political frameworks designed to adapt to changing relationships between society and ecosystems, institutional frameworks that enable adaptive management, and the facilitation of learning from adaptive management to policy.

Adaptive management A systematic process of natural resource management whereby management actions are treated as experiments to increase learning and improve subsequent management.

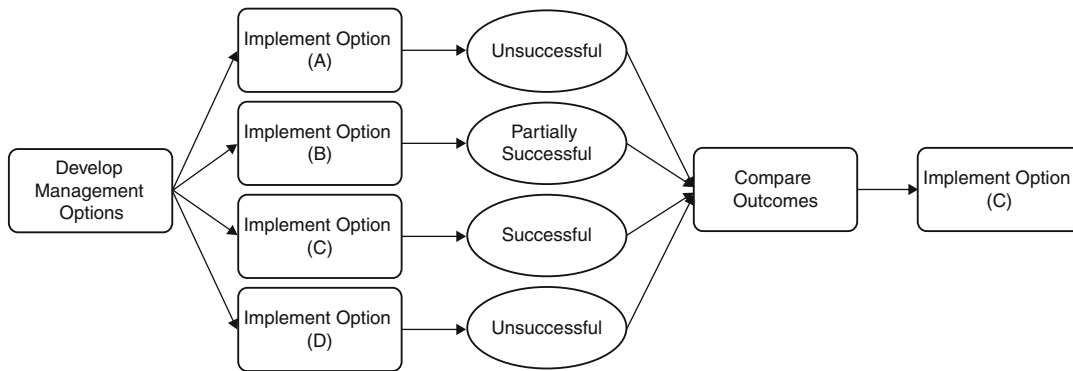
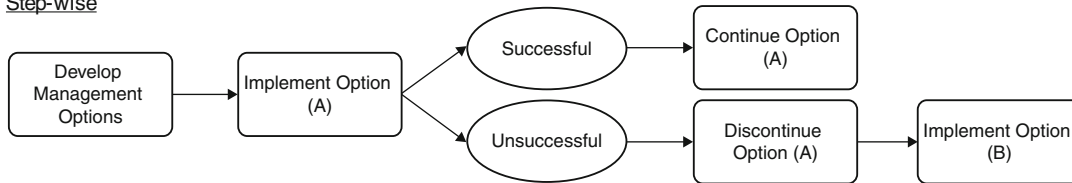
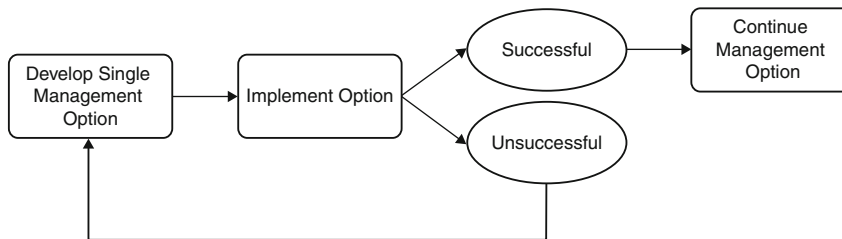
Natural resource management The management of natural resources including land, water, plants, and animals to meet societal goals, including conservation and exploitation.

Resilience The capacity of a system to absorb disturbance without altering states (undergoing a regime shift); a measure of the amount of disturbance a system can tolerate before collapsing.

Structured decision making A general term for a framework of analysis of problems to reach decisions based on evidence to meet stated goals.

Definition of Adaptive Management

Adaptive management is an approach to natural resource management that emphasizes learning through management based upon the philosophy that knowledge is incomplete and much of what is thought to be known is actually wrong, but despite uncertainty, managers and policymakers must act [1]. Although the concept of adaptive management has resonated with resource management scientists and practitioners following its formal introduction in 1978 [2], it has and continues to remain little practiced and much misunderstood. Misunderstanding is largely based upon the belief that adaptive management is what management has always been, a trial and error attempt to improve management outcomes. But unlike a trial and error approach, adaptive management has explicit structure, including a careful elucidation of goals, identification of alternative management objectives and hypotheses of causation, and procedures for the collection of data followed by evaluation and reiteration. Since its initial introduction and description, adaptive management has been hailed as a solution to endless trial and error approaches to complex natural resource management challenges and recently, it has become increasingly referenced under various forms (please refer to following sections) (Fig. 1). Regardless of the particular definition of adaptive management used, and there are many, adaptive management emphasizes learning and subsequent adaptation of management based upon that learning. The process is iterative, and

Horse RaceStep-wiseTrial and ErrorUncorroborated

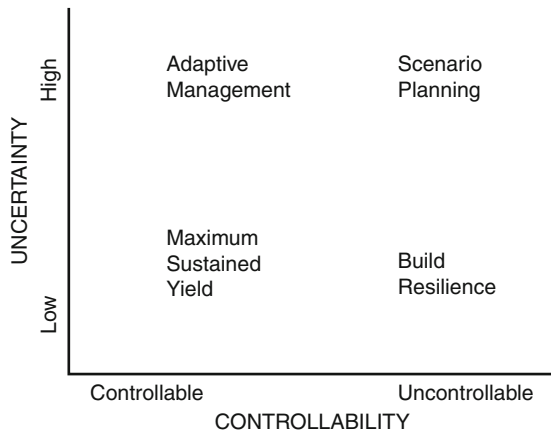
E

Ecosystems, Adaptive Management. Figure 1

Generalization of the different approaches to natural resource management

serves to reduce uncertainty, build knowledge, and improve management over time in a goal-oriented and structured process. However, adaptive management is not a panacea for the navigation of “wicked problems” [3, 4] as it does not produce easy answers, and is appropriate in only a subset of natural resource management problems where both uncertainty and

controllability are high (Fig. 2) [5]. Where uncertainty is high but controllability is low, scenarios are a more appropriate approach. Adaptive management is a poor fit for solving problems of intricate complexity, high external influences, long time spans, high structural uncertainty, and with low confidence in assessments [5] (e.g., climate change). However, even in such



Ecosystems, Adaptive Management. Figure 2

Adaptive management and scenarios are complementary approaches to understanding complex systems. Adaptive management functions best when both uncertainty and controllability are high, which means the potential for learning is high, and the system can be manipulated (Adapted from [60])

situations, adaptive management may be the preferred alternative, and can be utilized to resolve or reduce structural uncertainty.

Clearly, adaptive management has matured, but it has also reached a crossroads. Its application is now common to a variety of complex resource management issues, and while practitioners and scientists have developed adaptive management and structured decision-making techniques, and mathematicians have developed approaches to reducing the uncertainties encountered in resource management, there continues to be misapplication of the method, and misunderstanding of its purpose.

Introduction

Adaptive management of natural resources did not spontaneously appear, but represents an evolving approach to natural resource management in particular, and structured decision making in general. Founded in the decision approaches of other fields [6] including business [7], experimental science [8], systems theory [9], and industrial ecology [10], the first reference to adaptive management philosophies in natural resource management may be traced back to Beverton and Holt [11] in fisheries management,

though the term “adaptive management” was yet to be used (reviewed in [6]). The term “adaptive management” would not become a common vernacular until C.S. Holling, widely recognized as the “father” of adaptive management, produced his edited volume on the subject “Adaptive Environmental Assessment and Management” in 1978 [2]. The work was spawned by the experiences of Holling and colleagues at the University of British Columbia following from the development of resilience theory [12]. The concept of resilience, predicated upon the existence of more than one alternative stable state for ecosystems, had several ramifications. For one, it meant that managers should be very careful not to exceed a threshold that might change the state of the system being managed, and the location of those thresholds is unknown. Second, for ecological systems in a favorable state, management should focus on maintaining that state, and its resilience. Adaptive management, then, was a method to probe the dynamics and resilience of systems while continuing with “management,” whereby management experiments were developed to enhance learning and reduce uncertainty, in a fail-safe manner. According to Holling (<http://www.resalliance.org/2561.php>):

- The resilience research led us to mobilize a series of studies of large scale ecosystems subject to management- terrestrial, fresh water and marine. All this was done with the key scientists and, in some cases, policy people who “owned” the systems and the data. So the process encouraged two major advances. One advance developed a sequence of workshop techniques so that we could work with experts to develop alternative explanatory models and suggestive policies. We learned an immense amount from the first experiment. That focused on the beautiful Gulf Islands, an archipelago off the coast of Vancouver. We chose to develop a recreational land simulation of recreational property. I knew little about speculation, but we made up a marvelous scheme that used the predation equations as the foundation- the land of various classes were the “prey,” speculators were the “predators” and a highest bidder auction cleared the market each year. The equations were modifications of the general predation equations. The predictions were astonishingly effective and persisted so for at least a decade. As much as anything, it reinforced the earlier conclusion that

these equations were powerful and general. But the important conclusion concerned the workshop process and the people.

Eventually Carl Walters [1] built upon Holling's foundational contribution [12] and further developed the ideas, especially in the realm of mathematical modeling. Whereas Holling's original emphasis was in bridging the gap between science and practice, Walters emphasized treating management activities as designed experiments meant to reduce uncertainty. Both scientists sought an approach that allowed resource management and exploitation to continue while explicitly embracing uncertainties and seeking to reduce them through management. Walters [1] described the process of adaptive management as beginning "with the central tenet that management involves a continual learning process that cannot conveniently be separated into functions like research and ongoing regulatory activities, and probably never converges to a state of blissful equilibrium involving full knowledge and optimum productivity." He characterized adaptive management as the process of defining and bounding the management problem, identifying and representing what is known through models of dynamics that identify assumptions and predictions so experience can further learning, identifying possible sources of uncertainty and identifying alternate hypotheses, and finally the design of policies to allow continued resource management or production while enhancing learning.

A key focus of adaptive management is the identification and reduction, where possible, of uncertainty. Uncertainty is reduced through management experiments which enhance learning. Williams [6] describes four critical sources of uncertainty:

1. *Environmental variation* is often the most common source of uncertainty, and is largely uncontrollable. It may have a dominating influence on natural resource systems, through such factors as random variability in climate.
2. *Partial observability* refers to uncertainty about resource status. An example of this is the sampling variation that arises in resource monitoring.
3. *Partial controllability* arises when indirect means (e.g., regulations) are used to implement an action (e.g., setting a harvest rate), and it can lead to the misrepresentation of management interventions

and thus to an inadequate accounting of their influence on resource behavior.

4. *Structural or process uncertainty* arises from a lack of understanding or agreement regarding the structure of biological and ecological relationships that drive resource dynamics.

Adaptive Management Today

Adaptive management has been referenced either implicitly [11] or explicitly [2, 13] for more than 50 years, but despite an illustrious theoretical history, there has remained imperfect realization of adaptive management in real world natural resource management decisions. The limited implementation of adaptive management stems from three fundamental problems: (1) a lack of clarity in definition and approach, (2) a paucity of success stories upon which to build [14–18], and (3) management, policy, and funding paradigms that favor reactive rather than proactive approaches to natural resource management [19, 20]. Each of these challenges has slowed the development of adaptive management as a paradigm for natural resource management and resulted in incomplete, inefficient, and even inappropriate implementation of adaptive management.

Although semantic arguments may seem the realm of ivory-towered professors, inconsistent and even contradictory approaches and definitions of adaptive management have resulted in confusion and limited the ability of management organizations to develop consistent and repeatable comprehensive adaptive management programs. Ironically, the confusion over the term "adaptive management" may stem from the flexibility inherent in the approach which has resulted in multiple interpretations of "adaptive management" that fall upon a continuum of complexity and *a priori* design, starting from the simple (e.g., "learning by doing") and progressing to the more explicit (e.g., "a rigorous process that should include sound planning and experimental design with a systematic evaluation process that links monitoring to management") [2, 21, 22]. Obviously, there is a clear distinction in intent, investment, and success between approaches that propose to learn from prior management decisions and those that outline a concise feedback mechanism dependent upon sound scientific principles on which

future management decisions will be made. The definition of “adaptive management” is further confused because one of the powerful attributes of adaptive management is the ability to simultaneously address multiple needs of managers, scientists, and stakeholders. The result has been published reports of adaptive management that emphasize definitions that focus on the needs of the authors and the ability of adaptive management to meet those needs (e.g., experimentation [14], uncertainty [23], changing management actions [24], monitoring [25], and stakeholder involvement [26]).

Despite the challenges in defining adaptive management, momentum and interest in the subject and its application continue to grow. The recent development by the United States Department of Interior of an adaptive management technical guide (<http://www.doi.gov/initiatives/AdaptiveManagement/TechGuide.pdf>) and the policies developed around this manual to:

- Incorporate adaptive management principles, as appropriate, into policies, plans, guidance, agreements, and other instruments for the management of resources under the Department’s jurisdiction. – Department of Interior Manual (522 DM 1)

are an indication of the growing movement in natural resource management toward taking a more proactive role in management decisions. Unfortunately, this movement has little to build upon with one clear exception, the U.S. Fish and Wildlife Service (USFWS) Adaptive Harvest Management Plan (AHM) for mid-continent mallards. Worldwide, AHM is one of the few successful efforts to apply the principles of adaptive management and demonstrate how to successfully manage natural resources by improving the understanding of natural systems through management actions. The adaptive management processes of AHM have greatly improved the understanding of the harvest potential of waterfowl populations, the ability of managers to regulate harvest, and the importance of monitoring and assessment programs to support the decision-making process.

So why has AHM succeeded while so many other attempts to implement adaptive management have stalled? First, AHM developed a clear and concise objective: maximize long-term waterfowl harvest while ensuring long-term viability of waterfowl

populations. The development and agreement by stakeholders to a concise set of fundamental objectives is paramount to ensuring the success of any adaptive management program. Failure to agree upon fundamental objectives and unwarranted attempts to alter objectives will ensure any attempt to manage, whether adaptive or not, will fail. The second key to the AHM success was due to simultaneous support for management, research, and monitoring. Waterfowl research and management in North America are nearly unequaled by almost any natural resource management program in terms of history, scope, and investment [27]. The enormity of historical and current data and the availability of resources for researchers and managers to utilize that data have facilitated the development of innumerable research and management activities all of which have fed back into the AHM process. In addition, the AHM program has arguably one of the most comprehensive monitoring programs for any ecological system currently under study. The combination of well-supported management, research, and monitoring programs has resulted in a clear reduction in the uncertainty of how waterfowl populations respond to management and enabled managers and policy makers to more effectively meet their stated objectives. Unfortunately, too often, attempts to implement adaptive management fail to address all of the requirements. In particular, resources for monitoring and research are often undervalued with the resultant outcome being a series of management actions with no understanding of their implications.

The final key to the success of AHM has been the ability to implement management and policy decisions based on the best information available. In many historical and current attempts to implement adaptive management, the regulatory body charged with implementation of management recommendations either is unable, or worse, is unwilling to implement actions proposed by the outcome of the adaptive management process. The body in charge of regulatory control is too often a stakeholder in the process of adaptive management with an agenda independent of regulating the resource alone. There may even be, and often are, several regulatory agencies controlling resources, each an independent stakeholder, each with an independent agenda. Such a situation can make implementation of a management

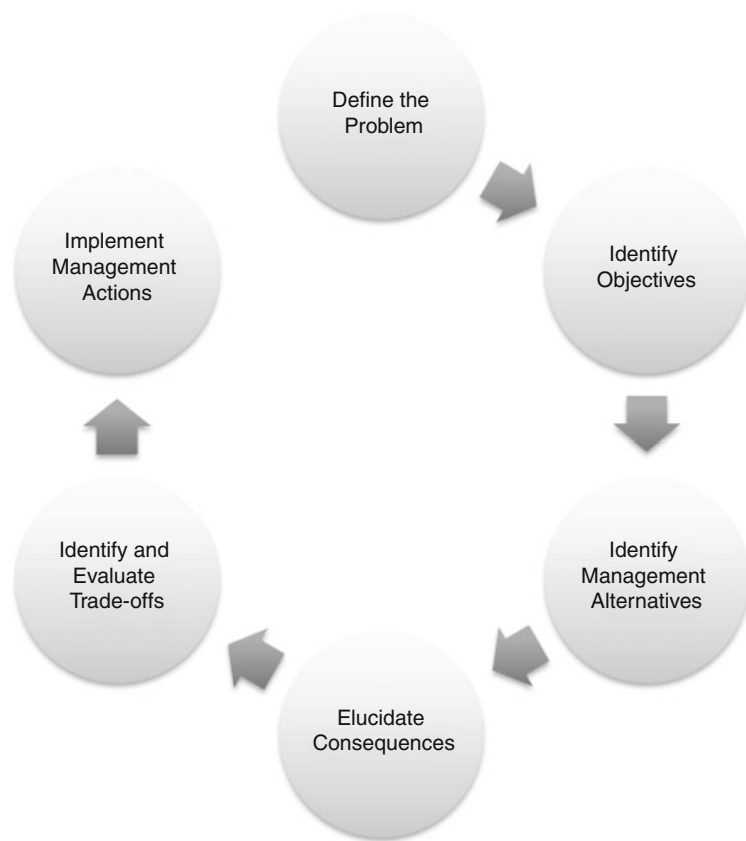
recommendation challenging, especially if it contradicts long-standing dogma. Consider for example, the management of Glen Canyon Dam and the waters of the Colorado River. Heralded by Congress as an adaptive management success story, the Colorado River Adaptive Management Program has fallen short of success because despite 13 years of work, the ecological status of the Colorado River and the conflict inherent to the development of an adaptive management program continue to worsen [28]. This is because the regulatory agency that controls the flow of water throughout the Colorado River Basin, the Bureau of Reclamation, is also one of the major stakeholders in the adaptive management process with an agenda (water storage) that conflicts with several other stakeholders and regulatory agencies that manage people and wildlife along the Colorado River (e.g., California Department of Water Resources, Mexican National Water Commission, USFWS). In contrast to the management of the Colorado River, there is a single centralized regulatory body governing waterfowl harvest in the United States (USFWS), and although there are many stakeholders that play a role in setting harvest management regulations, ultimately, decisions are made by the USFWS. Equally important, the interests of the USFWS parallel those of the other stakeholders. For the Colorado River, stakeholder interests are almost directly at odds. So from these examples is one to conclude that adaptive management is an unattainable mandate for the management of resources where various stakeholders and regulators are at odds? No, implementation of adaptive management is appropriate in both examples, possibly even more so for the management of the Colorado River. What the Colorado River example highlights is the importance of collaboration, the benefits of a single or superregulatory body, and the need to agree upon *a priori* objectives that guide long-term management decisions despite short-term political, societal, economic, or even environmental impacts.

Structured Decision Making

A key component of any management approach, whether it is adaptive or not, is deciding on the objectives, goals, and ultimately management options that may best achieve the desired goals (Fig. 3).

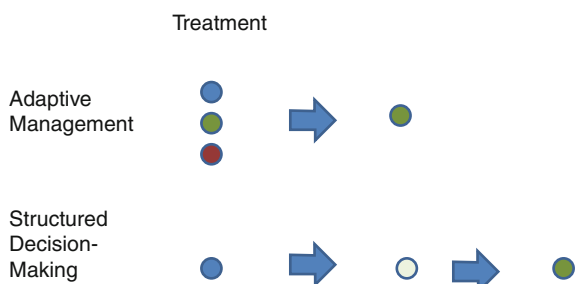
Unfortunately, as with many decisions, deciding upon a proper set of objectives and the means to reach those objectives can prove challenging. Resource management decisions are further complicated because social-ecological systems are complex (e.g., multiple objectives and stakeholders, overlapping jurisdictions, short- and long-term effects) and are characterized by a high degree of uncertainty (e.g., appropriate management action or monitoring protocols, future economic or ecological conditions) and therefore present decision makers with challenging judgments (e.g., predicted consequences of proposed alternatives, value-based judgments about priorities, preferences, and risk tolerances) often under enormous pressure (economic, environmental, social, and political) and with limited resources to ensure success. The resulting outcome of such conditions too often leads to management paralysis, or continuation of the status quo, as managers and policy makers become overwhelmed by the process of the decision and lose track of the desired social-ecological conditions they are charged with achieving. Indeed, the process of resource management can be arduous and even controversial, particularly if there are a variety of stakeholders vying to push the agenda. Fortunately, there are methods to overcome these pitfalls and maximize the potential for success.

One method to overcome management paralysis and mediate multiple stakeholder interests is structured decision making. Borrowed from the sociological fields, structured decision making is an organized approach to identify and evaluate alternative resource management options by engaging stakeholders, experts, and decision makers in the decision process and addressing the complexity and uncertainty inherent in resource management in a proactive and transparent manner. Structured decision making uses a simple set of steps (Fig. 3) to evaluate a problem and integrate planning, analysis, and management into a transparent process that provides a roadmap focused on achieving the fundamental objectives of the program. It differs somewhat from “active” adaptive management in that it does not emphasize replicated management experiments (Fig. 4). Central to the success of the structured decision making process is the requirement to clearly articulate fundamental objectives, explicitly acknowledge uncertainty, and respond



Ecosystems, Adaptive Management. Figure 3

The minimum steps necessary to implement a structured decision-making process: More complex integration of individual steps may be necessary if future steps clarify the process or if the decision is iterative over time



Ecosystems, Adaptive Management. Figure 4

Structured decision making and adaptive management differ somewhat, especially in that active adaptive management emphasizes the utilization of multiple replicated management experiments. As such, learning may be faster when such experiments are possible. However, adaptive management and structured decision making are terms often used interchangeably

transparently to all stakeholders’ interests in the decision process. The conceptual simplicity inherent in structured decision making makes the process useful for all decisions from minor decisions to complex problems involving multiple stakeholders.

Structured Decision Making Steps

1. *Define the Problem* – The first step in a structured decision making process is a clear and concise evaluation and articulation of the problem being addressed and the motivation underlying the need to address the problem. Although identifying the problem may seem self-evident, failure to clearly articulate the problem to all stakeholders and subsequent agreement by stakeholders as to the nature of the problem is often cited as the primary reason management and policy actions fail, or worse, face

future litigation. To facilitate this process, decision makers need to ask:

- (a) What specific decision(s) have to be made?
- (b) What is scope of the decision (e.g., geographic, temporal)?
- (c) Will the decision be iterated over time?
- (d) What are the constraints within which the decision will be made (e.g., logistical, ecological, legal, temporal, financial)?
- (e) What stakeholders should be involved in the decision process and what are their respective roles?

2. *Identify the Objectives* – The centerpiece of the structured decision making process is a set of clearly elucidated objectives. Together they define the “why do we care” about the decision and thereby facilitate the search for alternatives, and become the metric for comparing and evaluating management outcomes. When defining objectives, there are many considerations to ensure that decision makers can adequately evaluate alternatives. Ideally, objectives are stated in quantitative terms that relate to parameters that can be measured and thus evaluated. More importantly, objectives are meant to focus efforts on the importance of the decision in a consistent and transparent manner that exposes key trade-offs and uncertainties so decision makers can generate creative and proactive alternatives. Objectives should be complete, controllable, concise, measurable, and understandable [29]. To achieve this end requires “brainstorming” with stakeholders to identify what is important about the decision at hand. The outcome of such an effort may produce a wide variety and often extensive list of objectives that will need to be simplified to focus on things that matter and the direction they need to move (e.g., maximize deer harvest or minimize erosion). It is important to note, that unlike goals or targets, objectives do not have specific quantitative outcomes (e.g., 50% increase), but are meant to define the preferred ends and the direction of change to meet that ends.

Once a list of objectives has been defined, it is important to separate the objectives into fundamental objectives (which reflect the ultimate goals) and means objectives (which are ways of achieving the ends) to ensure that management

actions really effect the defined problem. For example, “maximize sandbars” may be an important objective for the management of a river like the Missouri or Platte, but if the river system is being managed for wildlife, sandbars are primarily important because they increase breeding habitat for threatened and endangered terns and plovers. “Maximize sandbars” is thus a means objective toward reaching the fundamental objective of “maximize tern and plover population size.” Clearly, there are other means objectives that would also facilitate this fundamental objective (e.g., minimize nest predation, maximized food availability, etc.). The benefit of the process of distinguishing objectives is that the identification of means objectives can help lead to alternative management actions (e.g., build sandbars, release reservoir water), while the identification of fundamental objectives gives a basis for evaluating and comparing alternatives (annual tern and plover population size). Keep in mind, however, that the status of fundamental or means is not an innate quality of an objective, but rather is highly context dependent. Thus, what was a means objective for one decision, in the example “maximize sandbars,” may be a fundamental objective for another if the decision problems shifts from say “wildlife management” to “aesthetics” or “flow.”

After developing a careful list of objectives, it can be useful to develop a hierarchy or means-ends diagram to group similar objectives and clarify the links and relationships between means and fundamental objectives. An objectives hierarchy can help clarify the context of each fundamental objective by identifying all the important elements that are affected by the decision process and demonstrate to stakeholders the importance of all objectives even those that are not “fundamental objectives.”

3. *Identify Management Alternatives* – Management success is only as likely as the creativity and diversity of possible management alternatives. Unfortunately, management paralysis, “pet” management actions, and staying with the status quo too often limit managers and policy makers to few options and thereby impede management success. The process of identifying management alternatives, like the process of identifying objectives, starts with

brainstorming. Identifying alternative management actions is a process that should be addressed iteratively, as knowledge of best practices and the creativity to develop novel ideas should not be expected to develop instantaneously. The key is bringing the “right” people together. It is important to have a group with a set of interdisciplinary backgrounds that represent the larger decision to ensure that the needs of stakeholders are not overlooked. This is not to say that the stakeholders involved in identifying alternative management actions are the same as the larger stakeholder group, usually they are not. This is primarily due to the technical knowledge necessary to present plausible alternatives. Still there are opportunities where the benefit of being naive may present novel actions that might not otherwise be considered.

The brainstorming process should begin by identifying alternatives for individual objectives, but always be looking for opportunities when one action may fulfill the needs of multiple objectives. Identifying alternatives also means being mindful of those actions that must be done (e.g., standing policy), constraints (real or perceived) and potential trade-offs between objectives and various management actions. In developing alternatives, it is important that the “brainstorming” process focus on developing high-quality management actions that are: (1) explicitly designed to address the outlined objectives, (2) technically sound in that they build on the best known practices, (3) concise yet comprehensive enough to include the technical understanding for implementation, (4) designed to expose trade-offs between the decision process by having mutually exclusive strategies, and (5) developed to achieve the greatest good for the stakeholders involved.

Once an extensive list of alternatives has been identified, it can be useful to group them into strategies or portfolios based on general similarities in what they aim to achieve. Sometimes these portfolios can represent the needs of specific stakeholder groups or specific conditions that could be achieved. For example, management actions on a river system may be grouped together into portfolios that meet the needs of sport-fishery, endangered species, or irrigation; alternatively, they may

be grouped based on their ability to return the river to 50%, 75%, or 95% of historical flows. Both methods have merit, the first in that it is generally clear to the stakeholders what objectives are being met and then where trade-offs must be considered, and the second in that the inherent interests of any particular group are not the driving factor and thus the process can be less contentious.

4. *Elucidate Consequences* – The list of alternative management actions is only effective if it creates an opportunity to evaluate and compare actions in light of the objectives before implementation. It is important to realize that the process of identifying management consequences is not a value judgment, but an analytical assessment of the most likely outcome of the action(s). Using the best scientific knowledge available, this process is a modeling exercise focused on predicting the likely outcomes of each alternative and thus the likelihood that each achieves the desired objective. Depending upon our knowledge of the system, this process can be highly quantitative where extensive data are modeled and probabilities assigned to each outcome or as is often the case, if little or nothing is known about the system, this process can depend heavily on expert opinion or comparisons to similar systems. In both cases, there is a degree of uncertainty associated with predicted outcomes as well as the parameters included in the modeling process. Indeed, because system function is rarely precisely understood, the effects of management actions are never certain and the future states are unknown, decisions are almost always made in the face of uncertainty. Uncertainty can make differentiating among alternatives difficult, but because uncertainty is an inherent part of the decision process, it must not be ignored. It is important that uncertainty be confronted throughout the decision process and that the uncertainties are identified and the possible impacts on the system and the ability to achieve stated objectives documented.

Once the modeling process has predicted the likely outcomes of each management action and the corresponding ability to address each objective, the next step is to develop a consequence table. The purpose of a consequence table is to produce

a visual summary of the consequences of each potential management action on each of the objectives in a table or matrix. A consequence table can take a variety of forms, from a simple rating system (e.g., consumer report 5-star rating) to a complex table with specific probabilities of outcomes and subsequent likelihoods of achieving each objective. Independent of the complexity of the underlying models that populate the matrix, the purpose of the consequence table is to ease and facilitate direct comparison of each management actions' ability to achieve each objective.

5. *Identify and Evaluate Tradeoffs* – Ideally the structured decision making process would lead to a clear management alternative that achieves the objectives of all interested parties; unfortunately, this is rarely the case. Generally, the process of developing a consequence table will clearly elucidate which options are the least likely to be effective, but if there are multiple stakeholders and thus multiple objectives, most decisions will require a trade-off between the ability of the remaining options to achieve each objective. The process of identifying where these trade-offs arise is analytical, but the decision process itself is highly value laden and thus dependent upon stakeholders. In most complex decisions, this will involve stakeholders choosing between less-than-perfect alternatives. There are a variety of methods to facilitate highly value-laden decisions by weighing options based on the values of the stakeholders and then comparing alternatives to find the “best” compromise solutions. However, trade-offs are real and it is unlikely that all parties will be totally satisfied with the eventual outcome. Indeed, although consensus is ideal, it is not necessary and is often unachievable; however, the benefit of the structured decision-making process is that even if there is disagreement, the process makes the disagreement transparent and enables stakeholders to re-evaluate using new knowledge and/or perspectives.
6. *Implement Management Action* – The final step in the structured decision-making process is implementation. Although this may always seem to be the desired outcome of a decision process, unfortunately, social and political pressures to reach “perfection” often impede implementation and

leave decisions in a continuous state of inaction. To ensure success, managers, policy makers, and stakeholders must work together to move through the decision process in a timely manner to ensure action can be taken. Failure to take action is a decision, whether it is made passively or actively.

Participatory Active Adaptive Management

Panarchy is a useful model for characterizing ecological systems and the formal institutions that manage these systems [30]. One of the most critical aspects in the panarchy appears to be a bridging organization that can monitor the status of the social-ecological system, and manifest rapid change, if conditions are deteriorating [31]. Monitoring will allow for management to set new target levels, and modify policy to reach those target levels, as new information is generated on scale-specific system attributes [32]. In order for management entities operating at discrete scales to improve communication channels and create opportunities for collaboration, intermediate level entities may serve to facilitate these cross-scale linkages. Bridging organizations have the capacity to fulfill this role and organize cooperation between stakeholders across scales [33], but to do so successfully, one must formulate strategies, coordinate joint action, address uncertainty, and link diverse stakeholders in a world of increasing complexity. Brown [33] investigated bridging organizations from across the world, and from a variety of scopes (e.g., regional economic policy in the USA; small-scale irrigation projects in Indonesia; agricultural productivity in Zimbabwe) found that bridging organizations are independent of stakeholders in a social-ecological system, which allows them to negotiate with stakeholders and advocate multiple positions. This unique role in the management of social-ecological systems affords bridging organizations the capacity to catalyze the formation of policies that are flexible and reflective of the panarchy of ecosystems and institutions [33]. In addition, bridging organizations have the capacity to reduce transaction costs, and provide a mechanism to enforce adherence to desired policies, despite their lack of regulatory authority [34].

Examples of bridging organizations include: (1) assessment teams, which are made up of actors across sectors in a social-ecological system;

(2) nongovernmental organizations, which create an arena for trust-building, learning, conflict resolution, and adaptive co-management; and (3) the scientific community, which acts as a “watchdog,” as well as a facilitator, for adaptive management. For purposes of environmental management, an example of a successful bridging organization is that of Ekomuseum Kristianstads Vattenrike (EKV), a small, municipal organization that facilitated progressive ecosystem management in southern Sweden [34]. EKV was tasked with managing water resources at a regional scale in Sweden, and was successful largely because it employed organizational flexibility that allowed for EKV to respond quickly to “surprise.” This was achieved through leadership, a core interdisciplinary staff, and the facilitation of connections between individuals and organizations (i.e., the panarchy of institutions) in the social-ecological system. EKV was able to improve the social capacity to respond to “surprises” and create the trust necessary to push the social-ecological system toward improved adaptive management of resources.

The formal management institutions in place are likely to persist barring a large-scale perturbation to social-ecological systems. So, managers must operate within the limitations of these institutions, which complicates matters, but does not make the situation intractable. One possible option for improving environmental management, as highlighted in this section, appears to be in developing bridging organizations that catalyze cross-scale communication across the panarchy of institutions and ecosystems, and explicit recognition of the underlying cross-scale structure and nonlinear interactions of these linked systems, by both policy and policy makers. The lack of communication and cooperation between institutions at even small scales further illuminates that bridging organizations may help bring about effective management of natural resources at multiple scales [35]. Thus, bridging organizations should act as mini think tanks that facilitate communication between institutions, incubate new ideas for environmental management, and provide a forum for coming to agreements on contentious issues [36].

Bridging organizations play a critical role in facilitating adaptive comanagement and governance, and are essential to managing for resilience in social-ecological systems [37]. Perception of a particular policy can play

a significant role in whether it is accepted by critical stakeholders in a social-ecological system [38]. Engaging stakeholders, implementing change at a suitable rate, and providing outreach to keep the public informed are all important for new environmental policy to be perceived of as positive and for a successful transition to a new policy regime [38]. This environmental management framework, which incorporates panarchy, adaptive management, and bridging organizations, could serve as one scenario in the suite of policy options for actualizing sustainability [30].

Adaptive Governance

Administrative agencies typically change incrementally [39], and as such, changes in policy are small because there is not enough information to make large overhauls of organization policy. Standard operating procedures are another mechanism that contributes to organizational inertia, as they slow the bureaucratic process [40]. Further, the lack of institutions matched to the appropriate scale is a significant barrier for sound environmental management [41]. Within this context, adaptive governance can help with this scale mismatch via collaboration of a diverse set of stakeholders at multiple scales [42]. Adaptive governance is a form of governance that incorporates formal institutions, informal groups/networks, and individuals at multiple scales for purposes of collaborative environmental management [43]. Bridging organizations, enabling legislation and government policies can also contribute to the success of an adaptive governance framework; governance creates a vision and management actualizes the vision [43].

Adaptive governance works via sharing of management power and responsibilities, and promotes a collaborative, participatory process, but is dependent upon adaptive comanagement, and adaptive comanagement is dependent upon social networks for success. Social networks have the capacity to allow for development of new ideas, to facilitate communication between entities, and to create the flexibility necessary for the interplay of the fluid (ecological systems) and the rigid (institutions) to be successful for environmental management [43]. Leadership has been well established as a critical factor in facilitating good environmental management. Leaders develop and facilitate a vision for

environmental management, incorporating local knowledge and information from social networks [43].

Olsson et al. [44] studied adaptive comanagement in Sweden and Canada and concluded that this form of management of ecological systems was most effective when there was: leadership with vision for the system of interest; legislation that created the environment for adaptive management; funds for adaptive management; monitoring of the ecological system; information flow (i.e., cross-scale linkages); combination of a variety of sources knowledge; and venue for collaboration. Olsson et al. [44] contend that these factors are critical to building resilience in social-ecological systems, as they help to protect the system from the failure of management decisions under uncertainty (i.e., imperfect information). Further, they assert that adaptive comanagement is necessary to facilitate adaptive governance. In turn, adaptive governance is facilitated by informal networks and leadership, which creates the capacity for development of novel ideas for environmental management [43]. These informal networks have the capacity to generate political, financial, and legal support for novel environmental management [43]. Further, adaptive governance is dependent upon polycentric institutions that are redundant (e.g., scale-specific) and are quasi-autonomous [45]. Olsson et al. [45] compared five case studies from around the world and concluded that in order for a social-ecological system to transition to adaptive governance, it must undergo a preparation and a transformation phase, linked by a window of opportunity.

In a well-cited example (Kristianstads Vattenrike) from Sweden, Olsson et al. [45] report the transition to adaptive governance was preceded by the development of a social network of parties interested in the management of the social-ecological system. The network consisted of members from local groups (environmental groups, farmers' associations), local government (municipality of Kristianstad, the County Administrative Board), and national scale (World Wildlife Fund, National Museum of Natural History, National Research Council). In case studies that have not resulted in a successful transition to adaptive governance, the social networks needed to help facilitate the transition were not well developed, and this hindered the changes needed for good environmental management [45].

The role of leadership has also been cited as critical to a transition to adaptive governance, and Olsson et al. [45] provide an example of leadership from Kristianstads Vattenrike. A key individual acted as a catalyst to social network formation, setting the research agenda, and mobilizing support at multiple scales for "new" environmental management. Critical to setting an agenda is defining how an issue becomes perceived as a "public problem because if most individuals accept a particular condition, negative feedback works to maintain public opinion in that particular regime" [46]. However, if the individuals in the regime develop a "critical mass" of distaste for a particular issue, public opinion can cross a threshold and reorganize into an alternative regime. Importantly, interest groups, the media, and other agents can have an effect on agenda setting and creating the "climate" necessary for a shift in public opinion [46]. There are critical roles to be played by individual actors in shifting policy from one regime to an alternate regime. For instance, social networkers that share information freely; individuals that have numerous, diverse connections; and individuals with powerful ability to persuade play key roles in policy change [47]. These individuals can interact to create the conditions necessary for regime shifts in public policy. In particular, the director of a municipal organization (Ekomuseum Kristianstads Vattenrike) filled this leadership role and served as a bridging organization that also was a significant factor in the transition to adaptive governance [45]. The leadership needed to foster a transition to adaptive governance is not necessarily the work of one individual, but rather is often encompassed by several individuals and entities [48].

There are two types of policy windows: a problem-driven window and a politically driven window [49]. A problem-driven window opens when a policymaker believes that a policy is necessary for a specific issue. A politically driven window is driven by a particular theme adopted by a policymaker, in which the policymaker looks for problems that fit within the theme. Significant changes in policy occur when conditions (e.g., problems, solutions, and politics) converge at the same time, which creates the window of opportunity for change [49]. In the Kristianstads Vattenrike example, social and ecological change at one scale triggered cross-scale effects which resulted in a window of opportunity for the transition to

adaptive governance [45]. In adaptive governance, decision making is not top-down but rather emerges from outreach and group meetings with stakeholders [50]. In order for adaptive governance to be effective, the policy requires strong leadership, communication, and incorporation of uncertainty, which allows for adaptation to changing circumstances [50].

Adaptive Management and Law

Legal certainty is an aspect of law that does not mesh well with environmental unpredictability. One of the most significant barriers for managing linked social-ecological systems is that often the aspects of a society that make it free (e.g., certainty of law) are not in concert with ecological realities (e.g., multi-regimes, nonlinear systems, and responses) [51]. The certainty of law and institutional rigidity often limit experimentation that is necessary for adaptive management [30]. This point is critical, as some scholars contend that environmental governance of the commons can only succeed if rules evolve with the system of interest [41].

Ecosystem management has been applied within the outdated framework of the Endangered Species Act (ESA), but ecosystem management is best implemented via adaptive management [52]. In its current form, the ESA does not have the necessary flexibility in its regulatory language to effectively implement adaptive responses to changing environmental conditions [52]. The legal constraints upon adaptive management in the American system of law do not stop there. The fundamental constraint to adaptive management is the current state of administrative law [53]. As the law now stands, the procedural rules require a vast amount of work before an agency promulgates a rule or issues a permit [54]. This “pre-decision” activity allows for public input and prepares agencies for judicial review. Ruhl [54] contends that “agencies will find that interest groups and courts relentlessly will erode adaptive agency behavior, using all the tools conventional administrative law puts at their disposal.” Having to operate in an atmosphere where each policy is evaluated on the “front-end,” in anticipation of public and legal scrutiny, has squelched agencies’ appetite for adaptive management.

US administrative law is a two-step process, in which the first step allows for public comment on

draft documents and alternative options [55]. The second step is final agency action, which creates “certainty” to the process and makes the decision subject to judicial review. This process is based on the assumption that agencies have the capacity to predict the consequences of a “final agency action” [55]. Thus, there is a fundamental conflict between linear legal processes (i.e., administrative law) based on “stationarity,” versus environmental management frameworks (i.e., adaptive management) based on the realization of dynamic systems characterized by “surprise” [55]. Given this inherent conflict, adaptive management may not be possible under the current administrative law framework [54].

The adversarial character of administrative law, combined with the need for certainty (e.g., procedural rules) in the larger realm of American law, is likely incompatible with adaptive management [56]. Thus, environmental law is at odds with science, as the certainty required for socio-political stability makes it very difficult to apply a novel approach to ecosystem management (e.g., adaptive management) that requires institutional flexibility. Thus, if adaptive management is necessary for good environmental management, environmental law must be “adapted” to fit with adaptive management [54]. Karkkainen [56] argues that administrative law should proceed on two trajectories: (1) a fixed rule track that will apply unless an agency can justify otherwise; and (2) an adaptive management track, where a new set of administrative law standards specific to adaptive management would hold precedence, in order to actualize adaptive management as a tool for environmental policy.

Thus, some in the law community argue that adaptive management is not possible under the current administrative law framework [54]. The National Environmental Policy Act (NEPA) may act as a barrier to implementation of adaptive management (*sensu* Holling) [57]. NEPA could possibly be modified to an iterative process that could accommodate adaptive management [57]. Ruhl [54] contends that adaptive management is necessary for good environmental management, which in turn means that environmental law must be “adapted” to fit with adaptive management.

In effect, administrative agencies in the USA do not conduct adaptive management as it was originally conceived [55]. Rather, agencies conduct adaptive

management “lite,” as the courts have provided some leeway to adaptive management projects, provided they have requirements that are legally enforceable [55]. The primary problem with adaptive management “lite” is that it does not measure up to the standards of adaptive management theory, nor does it hold up under the scrutiny of substantive and procedural law. Adaptive management (*sensu* Holling) is not likely until Congress provides more funding for adaptive management and clear standards for the adaptive management process [55].

Conclusions

The conceptual underpinnings for adaptive management are simple; there will always be inherent uncertainty and unpredictability in the dynamics and behavior of complex ecological systems as a result of nonlinear interactions among components and emergence, yet management decisions must still be made. The strength of adaptive management is in the recognition and confrontation of such uncertainty. Rather than ignore uncertainty, or use it to preclude management actions, adaptive management can foster resilience and flexibility to cope with an uncertain future, and develop safe-to-fail management approaches that acknowledge inevitable changes and surprises. Since its initial introduction, adaptive management has been hailed as a solution to endless trial and error approaches to complex natural resource management challenges. However, it does not produce easy answers, and it is appropriate in only a subset of natural resource management problems. Clearly adaptive management has great potential when applied appropriately.

Future Directions

Adaptive management is increasingly heralded as the future of natural resource management and has been adopted by many governmental and nongovernmental agencies. Institutions adopting adaptive management have utilized different definitions often focusing on a single strength of the process (*i.e.*, experimentation, reducing uncertainty, involving stakeholders) and thus operationalize the practice uniquely. Some, like the U.S. Department of Interior, are highly focused on the decision process and the incorporation of structured

decision making while others, such as the US Army Corps of Engineers, have embraced stakeholder involvement. Each approach has merit but adaptive management has failed to live up to its expectations [58]. The reasons for failure are many, and likely to be repeated, yet the great potential of adaptive management remains; unfortunately, it remains largely untapped. Translation of adaptive management approaches to “on-the-ground” natural resource managers is a critical step that has largely failed. Most natural resource managers are still unable to define adaptive management, let alone incorporate it into their normal management activities. The next decade will be critical: Will adaptive management remain in the domain of ivory towers, or will it become a tool for the trenches? Taking adaptive management to the practitioners will require the communication of adaptive management techniques in a clear, simple, and most importantly applicable manner. Currently, adaptive management fails because of an adherence to mathematical modeling above all else, its application to situations that are not conducive to replication or the measurement of success (*e.g.*, large rivers such as the Missouri or the Colorado), and because adaptive management has not been adequately incorporated into natural resources management via appropriate legal mechanisms [59]. If the future of natural resource management is to be proactive and address the increasing uncertainties facing our world, adaptive approaches to resource management will require communication of the methodology and merits in a clear and simple manner.

Acknowledgments

The Nebraska Cooperative Fish and Wildlife Research Unit is jointly supported by a cooperative agreement between the United States Geological Survey, the Nebraska Game and Parks Commission, the University of Nebraska—Lincoln, the United States Fish and Wildlife Service and the Wildlife Management Institute. Any use of trade names is for descriptive purposes only and does not imply endorsement by the US Government.

Bibliography

1. Walters CJ (1986) Adaptive management of renewable resources. McMillan, New York
2. Holling CS (ed) (1978) Adaptive environmental assessment and management. Wiley, Chichester

3. Rittel HWJ, Webber MM (1973) Dilemmas in a general theory of planning. *Policy Sci* 4:155–173
4. Ludwig D (2001) The era of management is over. *Ecosystems* 4:758–764
5. Gregory R, Ohlson D, Arvai J (2006) Deconstructing adaptive management: criteria for applications to environmental management. *Ecol Appl* 16:2411–2425
6. Williams BK (2010) Adaptive management of natural resources – framework and issues. *J Environ Manage* 92(5):1346–1353
7. Senge PM (1990) *The fifth discipline: the art and practice of the learning organization*. Currency Doubleday, New York
8. Popper KR (1968) *The logic of scientific discovery*, 2nd edn. Harper and Row, New York
9. Ashworth MJ (1982) *Feedback design of systems with significant uncertainty*. Research Studies Press, Chichester
10. Allenby BR, Richards DJ (1994) *The greening of industrial ecosystems*. National Academy Press, Washington, DC
11. Beverton RJH, Holt SJ (1957) *On the dynamics of exploited fish populations*. Her Majesty's Stationery Office, London
12. Holling CS (1973) Resilience and stability of ecological systems. *Annu Rev Ecol Syst* 4:1–23
13. Walters CJ, Hilborn R (1978) Ecological optimization and adaptive management. *Annu Rev Ecol Syst* 9:157–188
14. Lee KN (1993) *Compass and gyroscope: integrating science and politics for the environment*. Island Press, Washington, DC
15. McLain RJ, Lee RG (1996) Adaptive management: promises and pitfalls. *Environ Manage* 20:437–448
16. Lee KN (1999) Appraising adaptive management. *Conserv Ecol* 3(2):3. <http://www.consecol.org/vol3/iss2/art3/>
17. Moir WH, Block WM (2001) Adaptive management on public lands in the United States: commitment or rhetoric? *Environ Manage* 28:141–148
18. Walters C (2007) Is adaptive management helping to solve fisheries problems? *Ambio* 36:304–307
19. Ascher W (2001) Coping with complexity and organizational interests in natural resource management. *Ecosystems* 4:742–757
20. Schreiber ESG, Bearlsin AR, Nicol J, Todd CR (2004) Adaptive management: a synthesis of current understanding and effective application. *Ecol Manage Restor* 5:177–182
21. Wilhere GF (2002) Adaptive management in habitat conservation plans. *Conserv Biol* 16:20–29
22. Aldridge CL, Boyce MS, Baydack RK (2004) Adaptive management of prairie grouse: how do we get there? *Wildl Soc Bull* 32:92–103
23. Williams BK, Johnson FA (1995) Adaptive management and the regulation of waterfowl harvests. *Wildl Soc Bull* 23:430–436
24. Rauscher HM (1999) Ecosystem management decision support for federal forests in the United States: a review. *For Ecol Manage* 114:173–197
25. Bormann BT, Haynes RW, Martin JR (2007) Adaptive management of forest ecosystems: did some rubber hit the road. *Bioscience* 57:186–191
26. Norton BG (1995) Ecological integrity and social values: at what scale? *Ecosyst Health* 1:228–241
27. Hawkins AS, Hanson RC, Nelson HK, Reeves HM (eds) (1984) *Flyways: pioneering waterfowl management in North America*. U.S. Government Printing Office, Washington, DC
28. Susskind L, Camacho AE, Schenk T (2010) Collaborative planning and adaptive management in Glen Canyon: a cautionary tale. *Columbia J Environ Law* 35:1–54
29. McDaniels T (2000) Creating and using objectives for ecological risk assessment and management. *Environ Sci Policy* 3:299–304
30. Garmestani AS, Allen CR, Cabezas H (2009) Panarchy, adaptive management and governance: policy options for building resilience. *Neb Law Rev* 87:1036–1054
31. Kinzig A, Starrett D, Arrow K, Aniyar S, Bolin B, Dasgupta P, Ehrlich P, Folke C, Hanemann M, Heal G, Hoel M, Jansson A, Jansson BO, Kautsky N, Levin S, Lubchenco J, Maler KG, Pacala SW, Schneider SH, Siniscalco D, Walker B (2003) Coping with uncertainty: a call for a new science-policy forum. *Ambio* 32:330–335
32. Karkkainen BC (2002) Collaborative ecosystem governance: scale, complexity and dynamism. *Virginia Environ Law J* 21:189–244
33. Brown LD (1993) Development bridging organizations and strategic management for social change. *Adv Strat Manag* 9:381–405
34. Hahn T, Olsson P, Folke C, Johansson K (2006) Trust-building, knowledge generation and organizational innovations: the role of a bridging organization for adaptive co-management of a wetland landscape around Kristianstad, Sweden. *Hum Ecol* 34:573–592
35. Roy AH, Wenger SJ, Fletcher TD, Walsh CJ, Ladson AR, Shuster WD, Thurston HW, Brown RR (2008) Impediments and solutions to sustainable, watershed-scale urban stormwater management: lessons from Australia and the United States. *Environ Manage* 42:344–359
36. Brown LD, Khagram S, Moore MH, Frumkin P (2000) Globalization, NGOs, and multi-sectoral relations. In: Nye JS, Donohue JD (eds) *Governance in a globalizing world*. Brookings Institution Press, Washington, DC, pp 271–296
37. Olsson P, Folke C, Galaz V, Hahn T, Schultz L (2007) Enhancing the fit through adaptive co-management: creating and maintaining bridging functions for matching scales in the Kristianstads Vattenrike Biosphere Reserve Sweden. *Ecol Soc* 12(1):28. <http://www.ecologyandsociety.org/vol12/iss1/art28/>
38. Marshall NA (2007) Can policy perception influence social resilience to policy change? *Fish Res* 86:216–227
39. Lindblom C (1959) The science of muddling through. *Public Adm Rev* 19:79–88
40. Allison G (1969) Conceptual models and the Cuban missile crisis. *Am Political Sci Rev* 63:689–718
41. Dietz T, Ostrom E, Stern PC (2003) The struggle to govern the commons. *Science* 302:1907–1912

42. Hughes TP, Bellwood DR, Folke C, Steneck RS, Wilson J (2005) New paradigms for supporting the resilience of marine ecosystems. *Trends Ecol Evol* 20:380–386
43. Folke C, Hahn T, Olsson P, Norberg J (2005) Adaptive governance of social- ecological systems. *Annu Rev Environ Resour* 30:441–473
44. Olsson P, Folke C, Berkes F (2004) Adaptive co-management for building resilience in social-ecological systems. *Environ Manage* 34:75–90
45. Olsson P, Gunderson LH, Carpenter SR, Ryan P, Lebel L, Folke C, Holling CS (2006) Shooting the rapids: navigating transitions to adaptive governance of social-ecological systems. *Ecol Soc* 11(1):18. <http://www.ecologyandsociety.org/vol11/iss1/art18/>
46. Wood BD, Doan A (2003) The politics of problem definition: applying and testing threshold models. *Am J Political Sci* 47:640–653
47. Gladwell M (2000) *The tipping point – how little things can make a big difference*. Little, Brown, Boston
48. Zellmer S, Gunderson L (2009) Why resilience may not always be a good thing: lessons in ecosystem restoration from Glen Canyon and the Everglades. *Neb Law Rev* 87:893–947
49. Kingdon JW (1995) *Agendas, alternatives and public policies*. Harper Collins College, New York
50. Steelman TA, Tucker DW (2005) The Camino Real: to care for the land and serve the people. In: Brunner RD et al (eds) *Adaptive governance: integrating science, policy, and decision making*. Columbia University Press, New York, pp 91–130
51. Folke C, Pritchard L, Berkes F, Colding J, Svedin U (2007) The problem of fit between ecosystems and institutions: ten years later. *Ecol Soc* 12(1):30. <http://www.ecologyandsociety.org/vol12/iss1/art30/>
52. Ruhl JB (2004) Taking adaptive management seriously: a case study of the Endangered Species Act. *Univ Kans Law Rev* 52:1249–1284
53. Ruhl JB (1998) The Endangered Species Act and private property: a matter of timing and location. *Cornell J Law Public Policy* 8:37–54
54. Ruhl JB (2008) Adaptive management for natural resources – inevitable, impossible, or both? *Rocky Mountain Law Institute* 54:11.01.-06
55. Ruhl JB, Fischman R (2010) Adaptive management in the courts. FSU College of Law, Public Law Research Paper No. 411; Indiana Legal Studies Research Paper No. 154. SSRN: <http://ssrn.com/abstract=1542632>
56. Karkkainen BC (2005) Panarchy and adaptive change: around the loop and back again. *Minn J Law Sci Technol* 7:59–77
57. Benson MH (2009) Integrating adaptive management and oil and gas development: existing obstacles and opportunities for reform. *Environ Law Rep* 39:10962–10978
58. Allen CR, Gunderson LH (2011) Pathology and failure in the design and implementation of adaptive management. *J Environ Manage* 92:1379–1384
59. Benson MH, Garmestani AS (2011) Embracing panarchy, building resilience and integrating adaptive management through a rebirth of the National Environmental Policy Act. *J Environ Manage* 92:1420–1427
60. Peterson GD, Cumming GS, Carpenter SR (2003) Scenario planning: a tool for conservation in an uncertain world. *Conserv Biol* 17:358–366

Ecotones and Ecological Gradients

SALIT KARK

The Biodiversity Research Group, Department of Ecology, Evolution, and Behavior, The Hebrew University of Jerusalem, Jerusalem, Israel
ARC Centre for Environmental Decisions (CEED)
School of Biological Sciences, The University of Queensland, St Lucia, QLD, Australia

Article Outline

Glossary
Definition of the Subject
Introduction
Conceptual Framework
History of Research on Ecotones and Biodiversity
Approaches for Measuring Ecotones
Patterns of Biodiversity in Ecotones
Processes Shaping Biodiversity and Sustainability in Ecotones
Implications for Biodiversity, Conservation Planning, and Sustainability
Future Directions
Bibliography

Glossary

Beta-diversity Also termed species turnover, beta-diversity refers to the change in species as one moves between habitats, communities, or ecosystems.

Divergence-with-gene-flow model of speciation A model explaining the process of species formation (speciation) in which new species diverge in the face of gene flow; the movement of genes within a group that results from mating with immigrant individuals.

Ecotone A transition zone between two or more different ecological communities or regions.

Ecotone effect The pattern of increased species richness (number of species) and abundance in ecotones and the occurrence of unique ecotonal species.

Edge effect The effect of the juxtaposition of contrasting environments on an ecosystem.

Geographic information systems (GIS) A computer-based system for creating and managing spatial data and associated attributes. It enables the capture, storage, retrieval, analysis, and display of spatial (location-based) data.

Remote sensing The science and art of obtaining information about an object, area, or phenomenon through the analysis of data acquired by a device that is not in contact with the object, area, or phenomenon under investigation (e.g., via a satellite image).

Definition of the Subject

Ecotones are areas of transition between ecological communities, ecosystems, or ecological regions (such as Mediterranean and desert). Ecotones often occur along ecological gradients. Such gradients are created as a result of spatial shifts in elevation, climate, soil, and many other environmental factors. Ecotones commonly coincide with areas of sharp climatic transition along environmental gradients. They occur at multiple spatial scales, from continental-scale transitions between major biomes to small-scale ecotones where local vegetation communities and microhabitats coincide. They show a diversity of boundary types that range from natural boundaries (e.g., altitudinal, latitudinal transitions) to human-generated ecotones (e.g., forest clear-cut edges or urban ecotones). Ecotones have been studied in the past four decades in an ecological context and in recent years are receiving increasing attention in the context of biodiversity conservation. Various studies have shown that species richness and abundances tend to peak in ecotonal areas, though exceptions to these patterns occur. Ecotones are “natural laboratories” for studying a range of evolutionary processes, such as the process by which new species form, also termed speciation. Some researchers argue that ecotones deserve

high-conservation investment, potentially serving as speciation and biodiversity centers. Because ecotones are often small in size and relatively rich in biodiversity, conservation efforts in these areas may prove to be an efficient and cost-effective conservation strategy.

Introduction

While substantial research in the past decades has focused on distinct ecological regions, communities, and ecosystems, areas of transition between these regions have received much less attention. These areas, nevertheless, sustain rich and unique biodiversity and may have an important role in sustaining future biodiversity. However, studies have shown that species richness and abundances tend to peak in ecotonal areas, though exceptions to this pattern occur. This is likely because ecotonal areas hold species from two or more neighboring communities and possibly due to high spatial heterogeneity and high speciation rates in ecotonal areas. In addition, ecotones have been shown, in some cases, to hold ecotonal species and genotypes that are uniquely characteristic of the ecotone.

Several studies have found evidence that some species are diverging in ecotonal regions, suggesting that transitional areas may serve as speciation centers. Recent work has shown increased genetic and morphological diversity in ecotonal regions compared with neighboring areas. Ecotonal populations often coincide with peripheral populations along a species range. Some researchers suggest that because ecotones often hold marginal populations that depend on other parts of the range for the maintenance of genetic diversity, they do not deserve much conservation interest. Alternatively, in the face of global climate and environmental changes, it has also been argued that ecotones deserve special conservation investment, potentially being speciation and biodiversity hotspots that hold relatively resilient populations. This entry will provide a short overview of the history of ecotone research, recent novel findings in the area and the implications for future conservation, biodiversity, and sustainability research.

Conceptual Framework

Ecotones are areas where ecological communities, ecosystems, or biotic regions coincide. They often

occur in areas of steep environmental transition, along environmental gradients. In these transitional regions, the environment rapidly shifts from one type to another based on abiotic (e.g., climatic) and/or biotic (e.g., community structure) factors [13, 19]. The origin of the word “ecotone” is in the Greek roots “oikos” (home) and “tonus” (tension). Many different definitions and terms have been used in the literature to describe ecotones and areas of ecological transition. These include boundary regions, borders, meeting zones, transitional zones, tension zones, zones of intermingling, and zones of transgression [19]. The definition often contrasts ecotones with more homogenous areas found on both sides of the transition or to the landscape as a whole. Ecotones can occur in both terrestrial and aquatic systems, and cover several spatial scales, from large spatial-scale ecotones, where biomes meet [40] to local-scale transitions, such as mountain treelines ([11]; see Fig. 1). These areas are sometimes considered to be dynamic zones of interaction between communities, which are unstable over time [19]. As suggested by Odum [30], Ecotones do not simply represent a boundary or an edge; the concept of an ecotone assumes the existence of active interaction between two or more ecosystems with properties that do not exist in either of the adjacent ecosystems. Ecotonal regions show a diversity of transitions that range from natural transitions (e.g., altitudinal and latitudinal transitions) to human-generated ecotones, sometimes termed anthropogenic ecotones (e.g., forest clear-cut edges or urban ecotones) [24], as shown in Fig. 1.

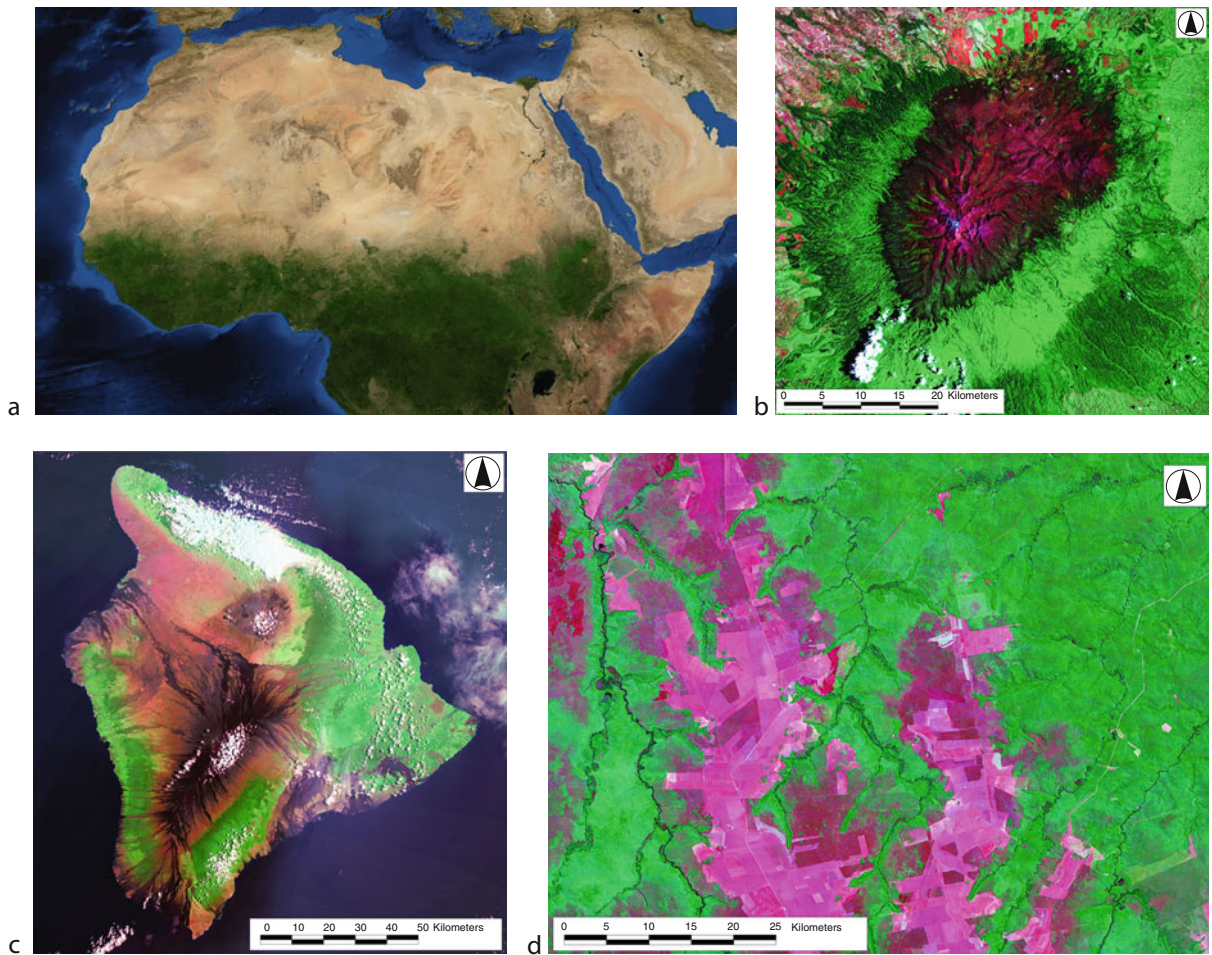
History of Research on Ecotones and Biodiversity

For over 100 years, since the late nineteenth century, there has been scientific interest in boundary regions and edges of ecological systems. A conceptual ecological framework for the study of ecotones was given by Odum in 1953. Odum discussed the ecotone effect, which he characterized as increased richness and abundance in ecotones and the occurrence of unique ecotonal species. Until the 1970s, there was considerable interest in ecotones within the scientific community (reviewed by [32]). This interest subsided as focus on more homogenous and well-defined ecosystems and communities (e.g., tropical rain forests and tundra) became common. A revival of research in the field

focusing on ecotones and their effects on biodiversity was seen in the late 1980s and 1990s, with the development of new research areas, especially those of landscape ecology and of conservation biology [11, 13]. Studies on ecotones in the 1980s often focused on material flow (e.g., water and nutrients) across communities and on ecosystem processes in these boundary regions [12]. Much of the work focused on wetlands and on riparian zones, where land–water interfaces occur (reviewed in [45]). Later work in the 1990s more directly examined the effect of ecotones on biological diversity, and especially on the relationship between ecotones and processes leading to morphological divergence, patterns of genetic and phenotypic diversity, species richness, rarity, and their conservation implications (reviewed in [17]).

Approaches for Measuring Ecotones

Due to the fact that ecotones can be rarely delimited by a fine line, their measurement and mapping is not simple. A wide range of research approaches and tools have been used to detect and quantify ecotones. These include, among others, simulation modeling, geographic information systems (GIS), remote sensing, and statistical tools that enable quantification and analysis of ecotones of different types and over several spatial scales. Diverse approaches for the quantification of the steepness of gradients exist [17]. Methods for measuring and characterizing ecotones depend on the data available (e.g., quantitative or qualitative, grid- or transect-based data), with one of the simplest approaches, proposed by Womble in 1951, being the quantification of the magnitude of the first and second derivatives (rates of change in a given variable or several variables along a spatial gradient) [44]. These approaches often examine the values of variables in an area (e.g., a 1×1 km grid square) relative to its neighboring regions. The basic idea is to detect areas of sharp environmental transition by finding the areas with the highest rate of change in the value of a given variable or several variables between adjacent squares (pixels). Specific software for the detection of boundary regions and analysis is now available (e.g., BoundarySeer: http://www.terraseer.com/products_boundaryseer.php), enabling more widespread use of advanced statistical tools for the study of areas of



Ecotones and Ecological Gradients. Figure 1

Satellite images showing a variety of natural and human-generated ecotones at several spatial scales: **(a)** Ecotones in the Sahel region of Africa (see text). **(b)** Altitudinal ecotones between vegetation belts in Mt. Kenya. **(c)** Natural and human-made ecotones on Hawaii's big island. Ecotones vary over different slopes and elevations. Note also the sharper ecotones in areas where lava has flowed. **(d)** Human-related ecotones resulting from deforestation (in pink) in the Amazon Basin, Brazil. On the far left and far right, in dark green, natural riparian ecotones can be detected along the rivers. (Source of Fig. **a** Stöckli et al. [39] NASA Blue Marble Next Generation, August 2004. True color composite of MODIS satellite images. Downloaded from: <http://visibleearth.nasa.gov/> [42].) (Source of Fig. **b–d** MDA Federal [25] NASA GeoCover circa 1990 Landsat images. False color composite of Landsat bands 7, 4 and 2. Green shades represent vegetation, magenta and brown represent bare soil, black represent water bodies, recent lava flows of shadows, cyan shades represent snow, and white represents clouds. Downloaded from: <https://zulu.ssc.nasa.gov/mrsid/> [3].) (All images are public domain. The figures were produced by Noam Levin)

transition [5, 8]. These tools can also be applied to the study of ecotonal regions.

In recent years, new approaches to quantify changes in diversity across gradients and boundary regions have been developed and are being applied. Among these is

a range of new beta-diversity estimates of species turnover in space [20, 26]. These have been developed in the past decades, since Wilson and Shmida's [43] review on beta-diversity estimates. Beta-diversity and species turnover are often used when studying gradients, and

although they do not focus necessarily on ecotonal areas, they can be applied to the study of ecotones.

One of the most promising directions in ecotone and boundary measurement is the use of tools developed in other areas of science. These include fields such as physics, remote sensing, and image analysis, where substantial advancements in boundary detection and gradient quantification methods have been made. Remote-sensing tools, for example, can use data occurring over several orders of magnitude, using satellite-derived data currently available at a resolution of 0.5 m to 100 km. Further application of these tools at multiple spatial resolutions will provide a better understanding of ecotones. Recent work has shown that remote sensing tools can be effectively used to detect ecotones and to predict species richness and rarity (e.g., range size rarity) in ecotones, especially in mountains where latitudinal gradients occur (e.g., [22, 23]).

Patterns of Biodiversity in Ecotones

There have been multiple studies on patterns of biodiversity in ecotonal areas, which have led to a range of results. Recent work is providing increasing evidence that boundary regions between ecological communities can be highly diverse at both the within-species and community levels. Ecotones have been shown to hold especially high biological diversity over several spatial scales, at both the community level (when examining species richness, i.e., the number of species in an area) and at the within-species level (morphological and genetic diversity) as reviewed by Kark and van Rensburg [17]. Other studies, however, have shown conflicting results, making it difficult to generalize without carefully examining each case, community, and region.

Early on, Odum [30] pointed at high species richness and abundance in ecotones and suggested that the ecotonal community commonly contains many species that are characteristic of, and sometimes restricted to, the ecotone. In a continental-scale study of New World birds, Kark et al. [16] examined the relationship between passerine richness and rarity of ~2,300 passeriform species in 4,889 one-degree New World grid cells and the distance of the cells to boundaries between adjacent plant-based ecoregions. They found that areas nearer to transitions between ecoregions had

more bird species, and also scored more highly in terms of species rarity. The findings of their work suggest that transitional environments harbor many rare species, in addition to high richness. At the community level, there is also some evidence for high species richness in ecotonal areas in marine systems. For example, van Rensburg et al. [41] showed at the subcontinental scale in South Africa that species richness and range size rarity at a spatial resolution of quarter degree are generally negatively correlated with distance to areas of transition areas between vegetation communities for both birds and frogs. Areas with more range-restricted species were located significantly closer to areas of transition areas between vegetation communities than expected by chance [41]. Similarly, in the Gulf of Aden, Kemp and colleagues [18] found high-reef fish diversity in an ecotone harboring a unique mixing of the three distinct faunas of Oman, the Red Sea, and the Indian Ocean.

Processes Shaping Biodiversity and Sustainability in Ecotones

Ecotones and Evolutionary Processes

Ecotones are “natural laboratories” for studying a range of evolutionary processes, such as the process by which new species form, also termed speciation. This process is of major interest to evolutionary biologists who define three major types of speciation: allopatric, parapatric, and sympatric. These models are based on the degree of geographical subdivision between populations that lead to the formation of new species. Allopatric speciation happens in geographical isolation, and has been for many years considered the major form of speciation. Parapatric speciation occurs in adjacent populations with gene flow among them, often along clines. Sympatric speciation occurs when populations are geographically congruent, and are found in the same area. The study of ecotones has led to a better understanding of the potential importance of parapatric and sympatric speciation as mechanisms for speciation. Ecotones have been proposed by some authors to be centers of evolutionary novelty that maintain evolutionary process, and as regions where parapatric (or sympatric) speciation processes may take place [33]. As such, ecotones and areas of environmental transition have been suggested

as natural laboratories where evolutionary processes and barriers to gene flow can be examined [33].

A review by Moritz et al. [27] summarized the major models of evolutionary processes that promote diversification of rain forest faunas. They include the divergence-with-gene-flow model, which suggests that adaptive divergence caused by selection forces occurs across environmental gradients, leading to speciation even in the face of gene flow across ecotones. This means that speciation does not require isolation in cases where selection is strong enough to separate populations. This process is expected to occur especially where very different environments meet in the ecotone, for example, at the border between a dry and wet rain forest. This may mean that even when gene flow continues, strong selection pressures can lead to divergence.

Support for the divergence-with-gene-flow model comes from recent research examining divergence using molecular genetic, phenotypic, and experimental approaches [27, 33, 34, 37]. Smith et al. [37], studying the little greenbul (*Andropadus virens*), a passeriform bird in the rain forest–savannah ecotone region of Cameroon, found especially high morphological divergence in the ecotone. The authors proposed that when the ecotone is large enough, natural selection processes could be strong enough to generate morphological differences similar to those seen in reproductively isolated species even when high rates of gene flow occur. Their data support the divergence-with-gene-flow model of speciation [38], leading them to propose that ecotones may be integral to the production and maintenance of high biodiversity in tropical rain forests. Quantification of morphological and geographic distances in olive sunbird (*Nectarinia olivacea*) populations in West African forests and ecotones revealed similar divergence patterns [38].

Smith studied morphological divergence in another Central African species, the black-bellied seedcracker (*Pyrenestes ostrinus*). Seedcrackers show polymorphism in bill size. A megabilled morph was found in the ecotone that specializes on a very hard-seeded sedge found only in ecotonal areas [27]. This ecotonal megamorph was maintained in the population, despite high levels of gene flow with rain forest populations that had only smaller-billed morphs, owing to its selective advantage for feeding on the hard seeds.

Morphological divergence between habitats across an ecotone was also found in leaf-litter skinks (*Carlia rubigularis*) in the wet tropical rain forest of Australia [34]. Adult skinks occurring across sharp ecotones from open (wet sclerophyll) forests to adjacent rain forests showed large morphological and life history differences over short distances despite moderate to high levels of mitochondrial gene flow [34]. Populations occurring across the ecotone had larger size differences than populations located dozens to hundreds of kilometers away that were geographically isolated millions of years ago, suggesting that in this case, and perhaps in others that await research, speciation with gene flow may have importance.

If ecotones regions harbor unique and endemic species and alleles, this may provide support for the notion that these regions may also have importance in speciation. If this is the case, ecotonal regions are expected to contain a preponderance of recently derived species that are yet to expand their ranges (neoendemics). Fjeldså and Rahbek [7] suggested that more recently evolved species are concentrated in transitional ecotones surrounding the main central African rain forest. This reasoning is consistent with the finding that terrestrial ecotones sustain high morphological divergence, providing evidence that current speciation processes may indeed be taking place in these regions [27, 33, 37].

Ecotones, Ecological Gradients, and Species Ranges

One of the important questions that arise is what are the processes that lead to higher species richness in ecotonal regions compared with adjacent areas? Evolutionary processes, as discussed above, may be one explanation. Ecological factors may also have importance in shaping this pattern. Several authors, such as Gosz [11] and Risser [32], have suggested that transitional areas not only share the two types of environments of the habitats that coincide in the ecotone, but also have a unique ecotonal environment. Indeed, Odum [30] proposed that transition zones often support a unique community with characteristics additional to those of the communities that adjoin the ecotone, although also commenting this is by no means a universal phenomenon. Studies testing these predictions show mixed results, some pointing toward

the occurrence of ecotonal species, while others not finding evidence for species unique and highly abundant in ecotones. The inconsistency among studies is complicated by the fact that different species, systems, scales, and regions were used in different studies or due to methodological factors, such as differences in sampling and analysis approaches.

In addition, ecotones tend to shift in space and time over several spatial scales [11, 19], as a response to climatic variation, other environmental changes [6, 19, 29], and human activity [10]. Ecotones show high spatial and temporal heterogeneity, which may serve as important factors contributing to their high genetic and species diversity [32]. For example, multiple ecotones can be defined within and around the African Sahel (Fig. 1a), depending on the scale of interest and on the definitions used [1]. The different transitions (e.g., that between the Sahel and the desert to the north) experience shifts in time and space, showing high spatiotemporal variability. Another, simpler, process shaping this pattern is that ecotones, comprising meeting areas between adjoining communities, include a combination of species from two or more community types [32]. Ecotonal areas often comprise the edge of the range for species on both sides and are where many peripheral populations occur [17, 35].

An important question is whether populations occurring in ecotonal areas are viable populations that exist over time within ecotones, or are rather present temporarily due to the constant flow of individuals from other parts of their range into the ecotone areas, are not self-sustainable over time, and will disappear if this flow is stopped. Shmida and Wilson [35] proposed that the high number of species in transitional areas could be due to a process they called the mass effect, which is the flow of individuals from favorable to unfavorable areas. For species that reach the margins of their range at the ecotone, this effect may result in some individuals of a given species establishing in ecotonal areas where they cannot maintain viable populations, existing in sinks adjacent to larger source populations [35]. This may lead to increased species richness in ecotonal areas, which is maintained by constant immigration of individuals from more favorable environments. Some evidence for the existence of mass effects can be found in the literature, yet these effects seem to be rather weak, and

it is currently unclear whether they can indeed act as a major factor generating high diversity in ecotones [21].

The mass effect, however, does not predict the occurrence of unique or endemic ecotonal species or genotypes. If some species or genotypes are characteristic of an ecotone or occur at the ecotone in higher abundances than in the neighboring habitats, as proposed above, this could suggest that some ecotone populations are ecologically viable. Several recent studies have found peak genetic and morphological diversity within species in ecotone regions, with populations in these regions harboring unique and rare alleles not found elsewhere [17]. For example, a study examining allozyme (protein level) diversity within chukar partridge (*Alectoris chukar*) populations across a rainfall gradient from northern to southern Israel found that the highest levels of diversity occur in the sharp ecotone area between the Mediterranean region and the desert, in the northern Negev area. Populations in this area not only showed higher genetic diversity based on 32 loci (proteins) examined, but also had unique and rare alleles that did not occur elsewhere across the range [14]. The same species also had peak levels of morphological diversity in the ecotone region based on 35 traits and 23 ratios between traits [15]. More studies on genetic and species uniqueness in ecotonal areas are needed to enable generalizations.

Agricultural, Urban and Human-Related Ecotones

The early literature discussing ecotones largely dealt with natural ecotones that are generated by environmental factors such as soils, geology, and climate. More recent research is increasingly including human-related boundaries. Human activity is generating boundaries that did not exist before, changing their steepness and shifting ecotonal location. These boundaries include a diversity of ecotone types, such as forest clear-cut edges, margins between built-up and natural landscapes, and human-generated features, such as lakes and plantations. Research on these human-generated ecotones and their effects on biodiversity are related to a study area that is sometimes termed “countryside biogeography,” examining biodiversity in human-dominated landscapes.

Another recently developing research area that is relevant is that of urban ecology. Sharp, human-caused transitions may result from human activities such as

urbanization, land-use changes, agriculture, grazing, or burning (see Fig. 1). These boundaries occur at multiple spatial scales, ranging from local ecotones between agricultural plots, urban areas, and roads, and their neighboring native habitat, to large-scale ecotones such as shifting desert borders owing to desertification processes and river divergence (Fig. 1). These ecotones may be either static and fixed in space or dynamic and shifting in location over space and time.

A substantial amount of work in both natural and human-related landscapes has focused on what has been called the “edge effect.” This is the effect of the juxtaposition of contrasting environments on an ecosystem. It refers to how the local environment changes along some type of boundary, or edge and how biodiversity is affected by such edges. This idea is attributed by animal ecologists to Aldo Leopold and his 1933 book *Game Management*. It encompasses a wide range of both biotic and abiotic trends associated with boundaries between adjacent habitat types, natural or anthropogenic. Much of the reference to edge effects in the recent landscape ecology literature has been related to human-caused boundaries, and especially to boundaries between forest fragments and neighboring patches of habitat that have been cleared. Again, there are no clear-cut conclusions as to the effect of human-generated ecotones on biodiversity. The response largely depends on the type of edge and its history as well the species in focus.

Its conservation implications are therefore complex and deserve further scientific attention. A vast amount of research has been done on the effect of forest edges, especially the effect of human-made forest fragments and their edges on biodiversity. A review of edges [28] suggested there are many discrepancies in the literature, and a better understanding and search for general patterns requires a much more mechanistic approach to examine the processes underlying such patterns. In this context, there has been much effort to understand the effect of forest edges on predation, brood parasitism, and the breeding success of birds. Paton [31] found that in the majority of studies, nest success varied near edges with an increase in both depredation and brood parasitism rates. The most conclusive studies suggest that edge effects in birds usually occur within 50 m of an edge [31]. Since these reviews, multiple studies on dozens of species and regions have been conducted.

Implications for Biodiversity, Conservation Planning, and Sustainability

Whereas better understanding of the effects of ecotones on biodiversity is important for evolutionary and ecological pure science purposes, it also has many implications for conservation of biodiversity, sustainability, and practical decision-making. Some scientists have argued that because ecotones hold marginal populations at the edge of the range of many species, where abundances may be lower and populations more prone to local extinctions than other parts of the range, these ecotonal areas have low value for conservation as they will not persist over time, for example, if the surrounding environments become fragmented (e.g., [9]). There has been an interesting discussion in the literature in recent years regarding whether transitional areas are valuable for conservation (e.g., see Smith et al. [36] versus Brooks [4]). Substantial conservation attention has been given in recent years to the understanding and mapping of biodiversity patterns and the underlying processes, and toward predicting the effects of global change. Ecotone and boundary regions, where change, shifts, and variability occur naturally in both space and time, could serve as useful models for understanding, monitoring, and predicting the response of individuals, populations, and communities to changing environments [2, 6, 29]. In addition, while some scientists suggest that ecotone populations are more likely to be negatively affected by climate change [9], other researchers have argued that ecotonal areas sustain populations that are adapted to changing, fluctuating, and unstable environments and, as such, these populations may better be able to persist in the face of predicted change (see further discussion in [17]). This is especially relevant since climate shifts are expected, according to some studies, to be rapid and extreme in boundary regions between ecosystems [2]. By examining changes in ecotone locations over time, these areas may potentially serve as “early warning” indicators of global changes [6, 29]. However, the response depends on the spatial and temporal scales examined and may be a useful indicator mainly at global spatial scales and rather coarse timescales. Therefore, this area deserves further attention owing to the complexity of the factors shaping the location of ecotones in space and time.

Much research in the past decade has focused on prioritizing conservation efforts and determining what areas are most important and valuable for conservation. Two main approaches have been suggested: The first approach includes a search for biodiversity hotspots, or areas with especially high species richness, endemism, and rarity. A second approach has been to select areas that are complementary, and hold biodiversity not present in other areas. Ecotones may provide a unique opportunity to conserve both high species richness and high complementarity. Due to their relatively small size, it may prove a cost-effective strategy to further conserve ecotone regions and to explicitly include them in future systematic conservation planning, given they potentially provide a high return on investment having small area and high biodiversity. Undoubtedly, if these areas have the potential to maintain and to generate species richness as well as unique and novel species and forms, they deserve far greater research attention than they are currently receiving [36]. Nevertheless, conservation plans for ecotones should not be considered independent of their surrounding environments. Ecotonal regions are important for our understanding of evolutionary (such as speciation, divergence with gene flow, and adaptation) and ecological processes shaping biodiversity (e.g., response of populations to fluctuating environments). They may enable us to better predict the responses of populations to environmental change and to further identify previously neglected biodiversity hotspots valuable for biodiversity research and conservation.

Future Directions

Research on the effect of ecotones on biodiversity is an important future direction in the face of global change, including land-use and climate change. Because ecotones are areas where shifts in environmental conditions occur in space and time, they may serve as useful indicators of environmental change and of the response of ecosystems to shifts in climate. The degree of effectiveness of ecotones as early predictors of the impacts of change and the ways that ecological communities and systems respond to change may be an important future direction. Ecotones can sustain unique forms or species that are less abundant or do not occur elsewhere. In addition, studies suggest that ecotones are areas where

some populations are diverging to new species in the face of gene flow (across the ecotone). It is important to examine the generality of these patterns and processes in different regions, spatial and temporal scales, and groups.

Another future direction is the application of multi-disciplinary approaches from other fields of science in studying and quantifying ecotones. For example, much work has been done in physics on transitions between states (liquid–solid etc.), which could be applied to work on environmental transitions ecotones. More work on the interactions between native and alien species and the role of ecotones in these interactions should be done to examine the hypothesis that ecotones are barriers to invasion between communities. The importance of ecotones to the generation and conservation of biodiversity, especially in the face of global climate and other environmental changes is an area of potential research and conservation investment that should further be explored and studied. The role of ecotones in promoting sustainability and its emerging science is an open area, which should be further explored in future work.

Bibliography

Primary Literature

1. Agnew CT, Chappell A (2000) Drought in the Sahel. *Geo J* 48:299–311
2. Allen CD, Breshears DD (1998) Drought-induced shift of a forest–woodland ecotone: rapid landscape response to climate variation. *Proc Natl Acad Sci* 95:14839–14842
3. Applied Sciences Directorate <https://zulu.ssc.nasa.gov/mrsid/>. Accessed Nov 2006
4. Brooks TM (2001) Prioritizing hotspots, representing transitions. *Trends Ecol Evol* 16:673
5. Colwell RK (2006) EstimateS: statistical estimation of species richness and shared species from samples. pp User's guide and application
6. Crumley CL (1993) Analyzing historic ecotonal shifts. *Ecol Appl* 3:377–384
7. Fjeldsø J, Rahbek C (1998) Continent-wide conservation priorities and diversification processes. In: Mace G, Balmford A, Ginsberg J (eds) *Conservation in a changing world*. Cambridge University Press, Cambridge, pp 107–138
8. Fortin M-J, Olson RJ, Ferson S et al (2000) Issues related to the detection of boundaries. *Landsc Ecol* 15:453–466
9. Gaston KJ, Rodrigues ASL, van Rensburg BJ, Koleff P, Chown SL (2001) Complementary representation and zones of ecological transition. *Ecol Lett* 4:4–9

10. Gehrig Fasel J, Guisan A, Zimmermann N (2007) Tree line shifts in the Swiss Alps: climate change or land abandonment? *J Veg Sci* 18:571–582
11. Gosz JR (1993) Ecotone hierarchies. *Ecol Appl* 3:369–376
12. Hansen A, Di Castri F (1992) *Landscape boundaries: consequences for biotic diversity and ecological flows*. Springer, New York
13. Holland M, Risser P, Naiman R (1991) *Ecotones*. Chapman & Hall, New York
14. Kark S, Alkon PU, Safriel UN, Randi E (1999) Conservation priorities for chukar partridge in Israel based on genetic diversity across an ecological gradient. *Conserv Biol* 13:542–552
15. Kark S, Mukerji T, Safriel UN, Noy-Meir I, Nissani R, Darvasi A (2002) Peak phenotypic diversity in an ecotone unveiled in the chukar partridge by a novel estimator in a dependent sample (EDS). *J Anim Ecol* 71:1015–1029
16. Kark S, Allnutt TF, Levin N, Manne LL, Williams PH (2007) The role of transitional areas as avian biodiversity centres. *Glob Ecol Biogeogr* 16:187–196
17. Kark S, van Rensburg BJ (2006) Ecotones: marginal or central areas of transition? *Israel J Ecol Evol* 52:29–53
18. Kemp JM (2000) Zoogeography of the coral reef fishes of the north-eastern Gulf of Aden, with eight new records of reef fishes from Arabia. *Fauna Arab* 18:293–321
19. Kent M, Gill WJ, Weaver RE, Armitage RP (1997) Landscape and plant community boundaries in biogeography. *Prog Phys Geogr* 21:315–353
20. Koleff P, Gaston KJ, Lennon JJ (2003) Measuring beta diversity for presence–absence data. *J Anim Ecol* 72:367–382
21. Kunin WE (1998) Biodiversity at the edge: a test of the importance of spatial “mass effects” in the Rothamsted Park Grass experiments. *Proc Natl Acad Sci* 95:207–212
22. Levanoni O, Levin N, Turbé A, Pe’er G, Kark S (2011) Can we predict butterfly diversity along an elevation gradient from space? *Ecography*. doi:10.1111/j.1600-0587.2010.06460.x, In press
23. Levin N, Shmida A, Levanoni O, Tamari H, Kark S (2007) Predicting mountain plant richness and rarity from space using satellite derived vegetation indices. *Divers Distrib* 13:692–703
24. Lloyd KM, McQueen AAM, Lee BJ, Wilson RCB, Walker S, Wilson JB (2000) Evidence on ecotone concepts from switch, environmental and anthropogenic ecotones. *J Veg Sci* 11:903–910
25. MDA Federal (2004) *Landsat Geocover (1990/TM) Edition Mosaics; Tiles N-37-00 and S-37-00 (Mt. Kenya), N-14-15_20 (Hawaii), and S-21-10 (Amazon)*. US Geological Survey, Sioux Falls, South Dakota
26. Mena J, Vázquez Domínguez E (2005) Species turnover on elevational gradients in small rodents. *Glob Ecol Biogeogr* 14:539–547
27. Moritz C, Patton JL, Schneider CJ, Smith TB (2000) Diversification of rainforest faunas: an integrated molecular approach. *Annu Rev Ecol Syst* 31:533–563
28. Murcia C (1995) Edge effects in fragmented forests: implications for conservation. *Trends Ecol Evol* 10:58–62
29. Neilson RP (1993) Transient ecotone response to climate change: some conceptual and modelling approaches. *Ecol Appl* 3:385–395
30. Odum EP (1953) *Fundamentals of ecology*. W. B. Saunders, Philadelphia
31. Paton PWC (1994) The effect of edge on avian nest success: how strong is the evidence? *Conserv Biol* 8:17–26
32. Risser PG (1995) The status of the science examining ecotones. *Bioscience* 45:318–325
33. Schilthuizen M (2000) Ecotone: speciation-prone. *Trends Ecol Evol* 15:130–131
34. Schneider CJ, Smith TB, Larison B, Moritz C et al (1999) A test of alternative models of diversification in tropical rainforests: ecological gradients vs. Rainforest refugia. *Proc Natl Acad Sci* 96:13869–13873
35. Shmida A, Wilson MV (1985) Biological determinants of species diversity. *J Biogeogr* 12:1–20
36. Smith TB, Kark S, Schneider CJ, Wayne RK, Moritz C (2001) Biodiversity hotspots and beyond: the need for conserving environmental transitions. *Trends Ecol Evol* 16:431
37. Smith TB, Wayne RK, Girman DJ, Bruford MW (1997) A role for ecotones in generating rainforest biodiversity. *Science* 276:1855–1857
38. Smith TB, Wayne RK, Girman DJ, Bruford MW (2000) Evaluating the divergence-with-gene-flow model in natural populations: the importance of ecotones in rainforest speciation. In: Bermingham E, Dick CW, Moritz C (eds) *Rainforests past and future*. University of Chicago Press, Chicago, pp 148–165
39. Stöckli R, Vermote E, Saleous N, Simmon R, Herring D (2005) The blue marble next generation – a true color earth dataset including seasonal dynamics from MODIS. NASA Earth Observatory, Goddard Space Flight Center, Greenbelt
40. van Rensburg BJ, Koleff P, Gaston KJ, Chown SL (2004) Spatial congruence of ecological transition at the regional scale in South Africa. *J Biogeogr* 31:843–854
41. van Rensburg BJ, Levin N, Kark S (2009) Spatial congruence between ecotones and range-restricted species: implications for conservation biogeography at the sub-continental scale. *Divers Distrib* 15:379–389
42. Visible Earth <http://visibleearth.nasa.gov/>
43. Wilson MV, Shmida A (1984) Measuring beta diversity with presence–absence data. *J Ecol* 72:1055–1064
44. Womble WH (1951) Differential systematics. *Science* 114:315–322
45. Zalewski M, Schiemer F, Thorpe J (2001) Fish and land-inland water ecotones: overview and synthesis. *Ecohydrol Hydrobiol* 1:261–266

Books and Reviews

- Blake JG, Loiselle BA (2000) Diversity of birds along an elevational gradient in the Cordillera Central, Costa Rica. *Auk* 117:663–686
- Kark S (2007, updated 2011) Effects of ecotones on biodiversity. In: Levin S (ed) *Encyclopedia of biodiversity*. Elsevier, Oxford, p 1–10

- Kluge J, Kessler M, Dunn R (2006) What drives elevational patterns of diversity? A test of geometric constraints, climate and species pool effects for pteridophytes on an elevational gradient in Costa Rica. *Glob Ecol Biogeogr* 15:358–371
- Mac Nally R, Molyneux G, Thomson J, Lake P, Read J (2008) Variation in widths of riparian-zone vegetation of higher-elevation streams and implications for conservation management. *Plant Ecol* 198:89–100
- McCain C (2009) Global analysis of bird elevational diversity. *Glob Ecol Biogeogr* 18:346–360
- Odum E, Barrett G (1971) *Fundamentals of ecology*, 5th edn. W.B. Saunders, Philadelphia
- Oommen M, Shanker K (2005) Elevational species richness patterns emerge from multiple local mechanisms in Himalayan woody plants. *Ecology* 86:3039–3047
- Sanchez-Rodriguez J, Baz A (1995) The effects of elevation on the butterfly communities of a Mediterranean mountain, Sierra de Javalambre, central Spain. *J Lepid Soc* 49:192–207
- Senft A (2009) Species diversity patterns at ecotones MSc thesis, University of North Carolina, Chapel Hill
- Terborgh J (1971) Distribution on environmental gradients: theory and a preliminary interpretation of distributional patterns in the avifauna of the Cordillera Vicabamba, Peru. *Ecology* 52:23–40

Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of

YIMIN GAO

Department of Electrical and Computer Engineering,
Texas A&M University, College Station, TX, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Architectures of Electric Vehicles

Architectures of Hybrid Vehicles

Architectures of Fuel Cell Vehicles

Future Directions

Bibliography

Glossary

Architecture The mechanical and/or electrical structure that defines the energy flow routes in a vehicle.

Electric coupling The method by which two electric powers are added together.

Electric vehicle An electric vehicle is referred to as a vehicle that uses electric energy as its energy source, generally chemical batteries and/or ultracapacitors and electric motor as its power source.

Fuel cell vehicle A fuel cell vehicle is referred to as a vehicle that uses fuel cells and electric motor as its power source.

Hybrid electric vehicle A hybrid electric vehicle is referred to as a vehicle in which two energy sources and correspondingly two power sources are installed, which are selectively used to power the vehicle. The primary energy source and power source may be chemical fuels and internal combustion engine. The secondary energy source and power source may be chemical batteries and electric motor.

Mechanical coupling The method by which two mechanical powers are added together.

Operation mode Drive train operation status in which power sources are selectively used as the active power mover to propel the vehicle.

Definition of the Subject

The architecture design of an electric vehicle is to design its energy flow route from the energy source to the energy destination. Usually, chemical batteries are used as its energy source and electric motors are used to produce mechanical power to propel the vehicle. For satisfying vehicle performance, the batteries, electric motor, and transmission should be properly designed.

A hybrid electric vehicle uses two power sources to power the vehicle. There are many methods to connect the two power sources together, each of which has its special operation characteristics. For different vehicles that have different mission requirements and operation environments, a special architecture should be used to fully use its advantages and avoid its shortcomings.

Fuel cell vehicle is the vehicle which is powered by a fuel cell system. Compared with internal combustion engine, fuel cell is more efficient and cleaner. Architectural design of a fuel cell vehicle consists in designing the drive train structure and components to meet the vehicle performance requirements.

Introduction

Electric vehicle (EV) is referred to as a vehicle that employs electric energy storage as its energy source and electric machines as its power source. Electric energy is a multisource energy type which can be obtained from many primary energy sources, such as traditional fossil energy sources (coal, petroleum, and natural gas), nuclear energy, hydraulic power, bioenergy, solar energy, wind energy, etc. With continuous depletion of petroleum resources and more and more concerns about environmental issues, it has been well recognized that electric vehicle is one the most viable substitutes to the current petroleum-fueled vehicle. Electric vehicle also possesses other advantages of absence emissions, high efficiency, quiet and smooth operation, etc. With continuous progress on the technologies of chemical batteries, electric propulsion systems, and electronic control, electric vehicles are much closer to meeting a user's requirements than ever before.

Compared with petroleum fuels, the energy density of chemical batteries is much lower, which results in shorter driving range per battery charge. Long time battery charging also causes inconvenience to users. At present, the battery-powered electric vehicles cannot challenge the petroleum-powered vehicles in terms of performance and use convenience. Hybrid electric vehicles are the combination of conventional vehicles and battery-powered electric vehicles. They can adopt the advantages and avoid the shortcomings of both. It has been recognized that hybrid electric vehicles, especially plug-in hybrid vehicles, are the major

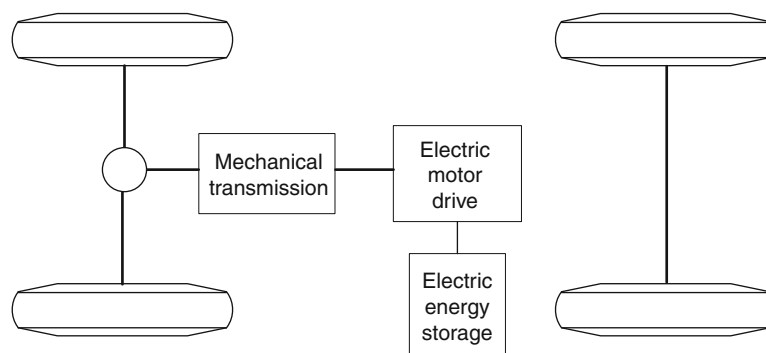
substitutes of conventional vehicles before technologies of chemical batteries have substantive progress in energy density, safety, and cost.

Fuel cell vehicles have been considered to be the final solution as the future high-efficiency and clean vehicles. Compared with conventional vehicles, fuel cell vehicles are independent from the petroleum resource and potential has high efficiency and zero emission. Compared with battery-powered electric vehicles, they potentially have the advantages of long driving range and quick fuel replenishing. However, at present, the hydrogen sources and onboard storage are the major challenges.

Architectures of Electric Vehicles

In principle, electric vehicles share the common characteristics of internal combustion engine-powered vehicles. A chemical battery pack replaces the fuel tank as its energy storage. Electric machine replaces the internal combustion engine as the power source. Both share the similar performance requirements, such as, speed, acceleration, gradeability, and driving range. The design principles as discussed in the chapter of [▶ Vehicle Dynamics and Performance](#) are still valid for electric vehicles. However, unique operating characteristics of the electric propulsion system have to be involved into the design process.

Previously, electric vehicles were designed to have a similar drive train structure to conventional internal combustion engine-powered vehicles as shown in [Fig. 1](#). It simply replaces the engine with an electric motor and fuel tank with a chemical battery pack. With



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 1
Primary electric vehicle drive train

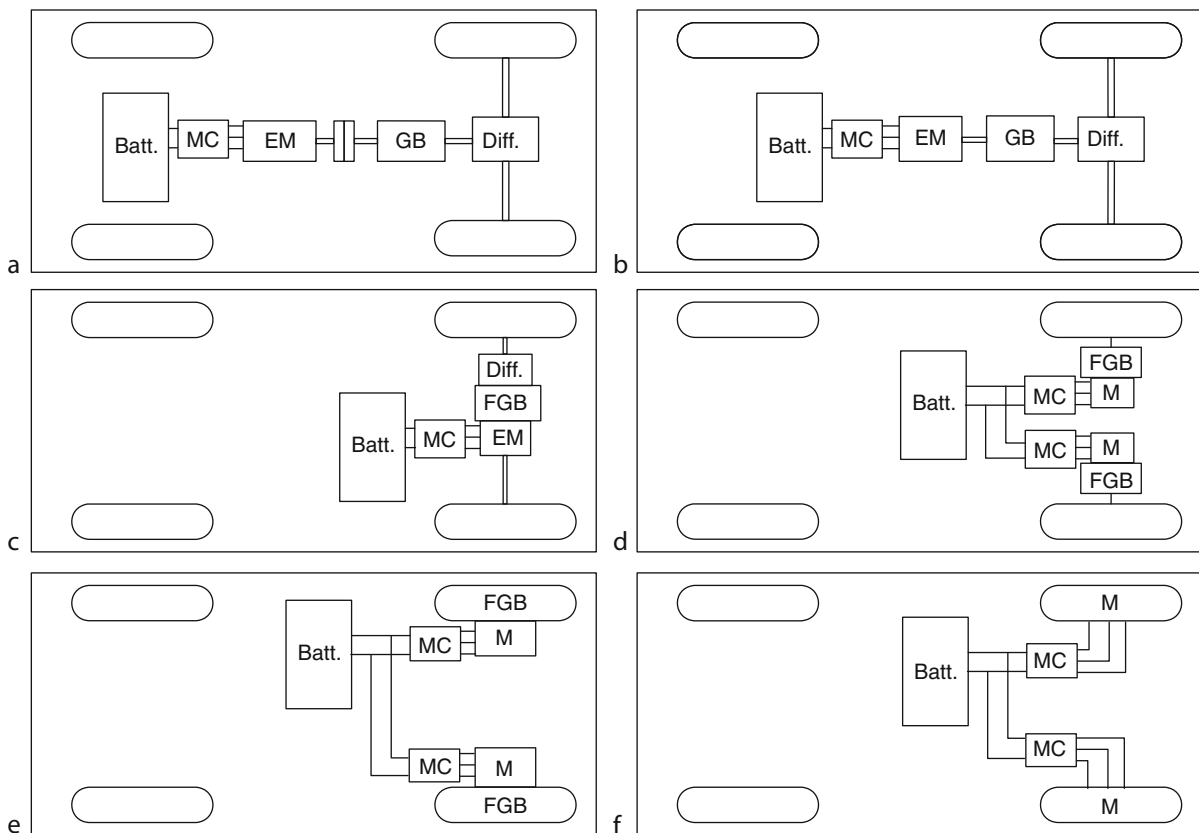
development of the technologies of light weight, compact electric drive system, and advanced control, an electric vehicle drive train can be designed to various configurations as shown in Fig. 2 [1].

Figure 2a shows the first alternative configuration that is most alike a conventional one. An electric motor (EM) drive replaces the internal combustion engine. A clutch is used to connect the motor drive to or disconnect from the final drive and differential. A transmission gear box (GB) is used to modify the output speed–torque profile for meeting the performance requirement and supply a reverse gear. This drive train may be used for the motor drives which are only capable of operating in one or two quadrants, forward motoring and/or forward regeneration. Since a transmission gearbox is used, the torque-speed profile may not be required to close the ideal profile, which has constant torque in low speed range and constant power in high speed range. A motor drive

with simple control may meet the requirement, such as series-excited DC motor.

Similar to the configuration shown in Fig. 2a, b shows the configuration in which clutch is not further needed. The motor starts producing its torque from zero speed. The transmission gear box may be either multi-gear or single gear, depending on the motor torque-speed profile. With a multi-gear transmission, the motor torque needs to be stopped during shifting. This can be carried out by cutting the current supply to the motor drive by the motor controller.

Figure 2c illustrates a configuration with a more compact design. A fixed gear ratio single gear transmission (FGB) is installed between the motor drive and the final drive (differential). This design requires the motor drive to be capable of producing a torque-speed profile which is closer to the ideal one. It also requires the motor be capable of operating in four quadrants, that is, forward and backward motoring and generating.



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 2
Possible electric vehicle configurations [1]

Figure 2d illustrates a configuration in which two motor drives are used. Each motor drives an individual wheel through a fixed gear ratio single gear box (FGB) which is used to match the motor speed to the wheel speed. A differential function is carried out by controlling the two motors. One advantage of this configuration is that it can fully utilize the road adhesive capability to improve the drivability of the vehicle. For example, when one wheel is on a slip road and loses its drive force, another wheel can produce a big driving force to help the vehicle get out of trouble.

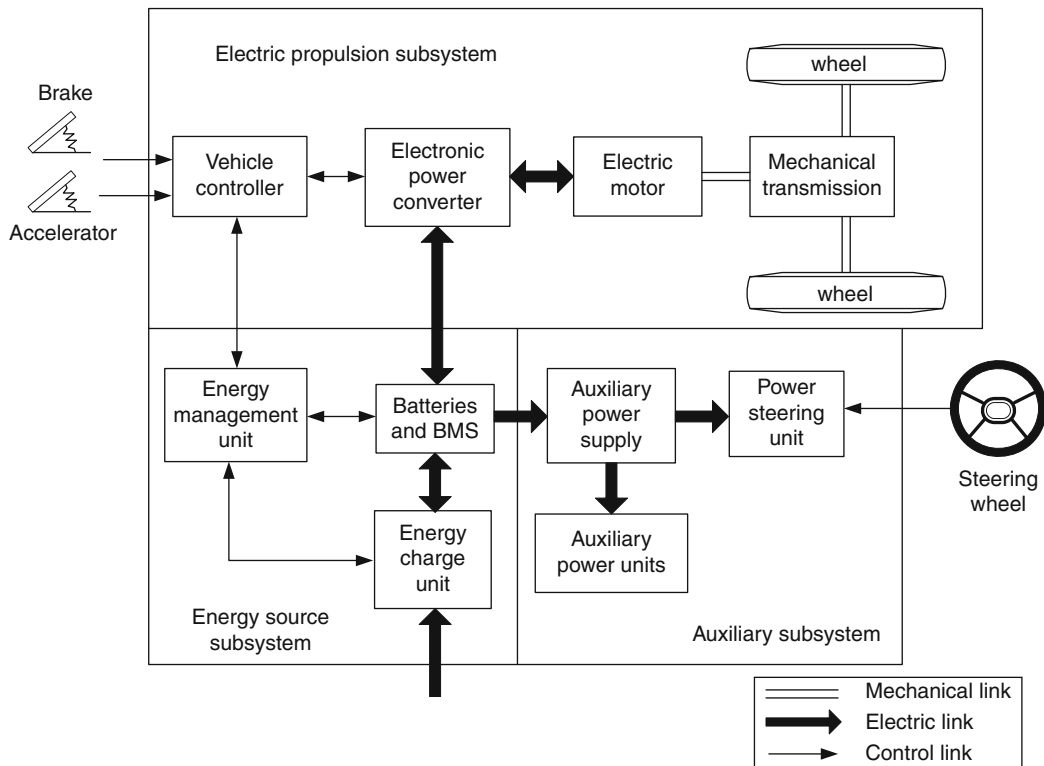
The drive train shown in Fig. 2e is very similar to the one in Fig. 2d. The difference is that the two fixed gear ratio single gear boxes (FGB) are attached to wheels rather than to the vehicle frame. This design can save much space and may make packaging easier.

Figure 2f shows a simplest drive train structure, in which the wheel is designed as the rotor of the motor. This configuration is the most compact one and saves a lot space. This structure is generally termed as

wheel motor. In the design of the drive trains shown in Fig. 2e, f, the impact of additional wheel weight to vehicle handling and vibration performance should be considered.

In all the architectures, the electric motors are powered by a battery pack and corresponding motor controller(s).

Except electric motor(s) and corresponding transmissions, a modern electric vehicle also needs other components for fulfilling its functions as shown in Fig. 3. The drive train consists of three major subsystems: electric propulsion subsystem, energy storage subsystem, and auxiliary subsystem. These three subsystems are interfaced through power routes and control signal routes. The electric energy from power grid is charged to the energy storage (batteries) through an energy charge system (battery charger). The energy storage delivers its energy to or absorbs energy from the electric motor through an electronic power converter which functions as a motor controller.



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 3
Conceptual illustration of an electric vehicle configuration [1]

The electric motor operation is controlled by a vehicle controller which commands the torque requirement to the motor controller based on the accelerator or brake pedal position and energy source operation parameters and energy management algorithm. The energy source also supplies energy to the vehicle auxiliaries. A battery management system (BMS) is used to monitor the operation of the battery pack, generate operation information, such as state-of-charge (SOC), and protect battery pack from abusing.

Architectures of Hybrid Vehicles

Hybrid vehicles (HV) are referred to as the vehicles that are powered by at least two energy sources and two power sources, in which at least one energy source and one power source have bidirectional energy flow capability, delivering energy to drive wheels and accepting energy from braking. When electric energy source and electric power source are used, they are referred to as hybrid electric vehicles (HEV). It is possible to replace the electric system with a hydraulic system. This kind of hybrid vehicle is referred to as hydraulic hybrid vehicles (HHV). Fuel cell vehicles (FCV) are usually referred to the vehicles in which fuel cells are used as the primary power source. A fuel cell vehicle may use two energy sources, fuel cells and chemical batteries, which is referred to as fuel cell hybrid vehicles (FCHV).

Since there are two power sources and only one energy sink (wheels), there are many ways to combine the two powers together. The design of the energy flow routes are referred to as drive train architecture design. Control of the energy flows in hybrid drive train is referred to as vehicle control strategy.

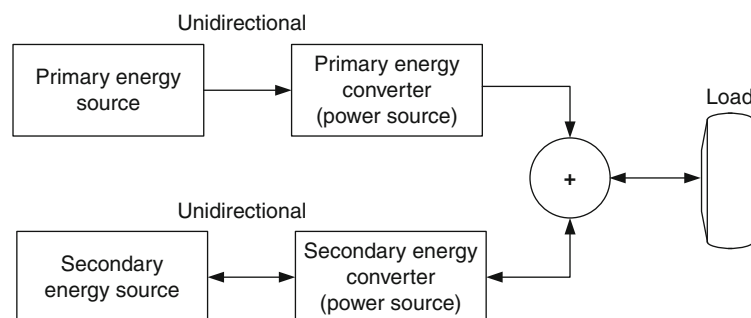
Concept of Hybrid Vehicle

In a hybrid vehicle, two power sources supply powers to drive wheels. The total power should meet the vehicle performance requirements, in terms of maximum speed, gradeability, acceleration time requirement, and driving range. The vehicle performance requires the vehicle possessing sufficient power, and sufficient energy for continuous driving range. In operation, any one of the two power sources or both can be used for traction. Thus, there are many possible operation modes as shown in Fig. 4 [2].

The primary energy source and energy converter allows unidirectional energy flow, whereas, the secondary energy source and energy converter are required to have the capability of bidirectional energy flow. With today's technologies, the common primary energy sources are chemical fuels and primary energy converters are internal combustion engines (ICE) and fuel cells. The secondary energy sources are chemical batteries and/or ultracapacitors and the secondary energy converters are electric motors. The final destination of the energy flow (energy sink) is the vehicle wheels with the form of mechanical energy.

As shown in Fig. 4, the wheels can be powered by selecting power sources, which are referred to as operation modes, which include:

1. Primary power source alone propelling
2. Secondary power source alone propelling
3. Both power sources propelling (hybrid drive)
4. Regenerative braking (secondary power source absorbs energy form load)



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 4
Concept of hybrid electric vehicles

5. Primary power source splitting for propelling and secondary power source charging

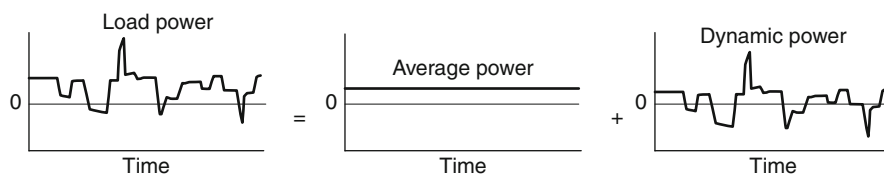
So many available operation modes create higher flexibility over single power source vehicles. It provides the drive train with opportunities for maximizing its operation efficiency and/or minimizing its emissions at any driving conditions by selectively using a best operation mode.

Vehicle load in normal driving varies randomly as shown in Fig. 5. Conceptually, the random load is composed of two components, one is the steady power (average) and the other is the dynamic power with zero average [2]. Ideally, hybrid vehicle can be designed and controlled in such a way that one power source supplies a constant power for the average load, which is referred to as energy source, since all the energy comes from it, and the other power source supplies dynamic power, which is named peaking power source, since it consumes zero energy in a whole driving cycle.

In normal driving, average power is relative small (few 10 kW). Thus, the primary power source can be designed with a small power capacity. When the internal combustion engine is used as a primary power source, it can be downsized and operated near its most efficient region. In addition, other types of engines may be used, such as Sterling engine, gas turbine engine, and others. The secondary power source may be batteries, ultracapacitors, flywheels and their combination, and electric motors.

Architectures of Hybrid Vehicles

Architecture of a hybrid vehicle is loosely defined as the ways by which the components of a hybrid drive train are connected and, therefore, the energy flow routes are schemed. Traditionally, hybrid vehicles are classified into series, parallel and series/parallel. However, the definitions of these configurations have not been clear.



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 5
Steady power and dynamic power [2]

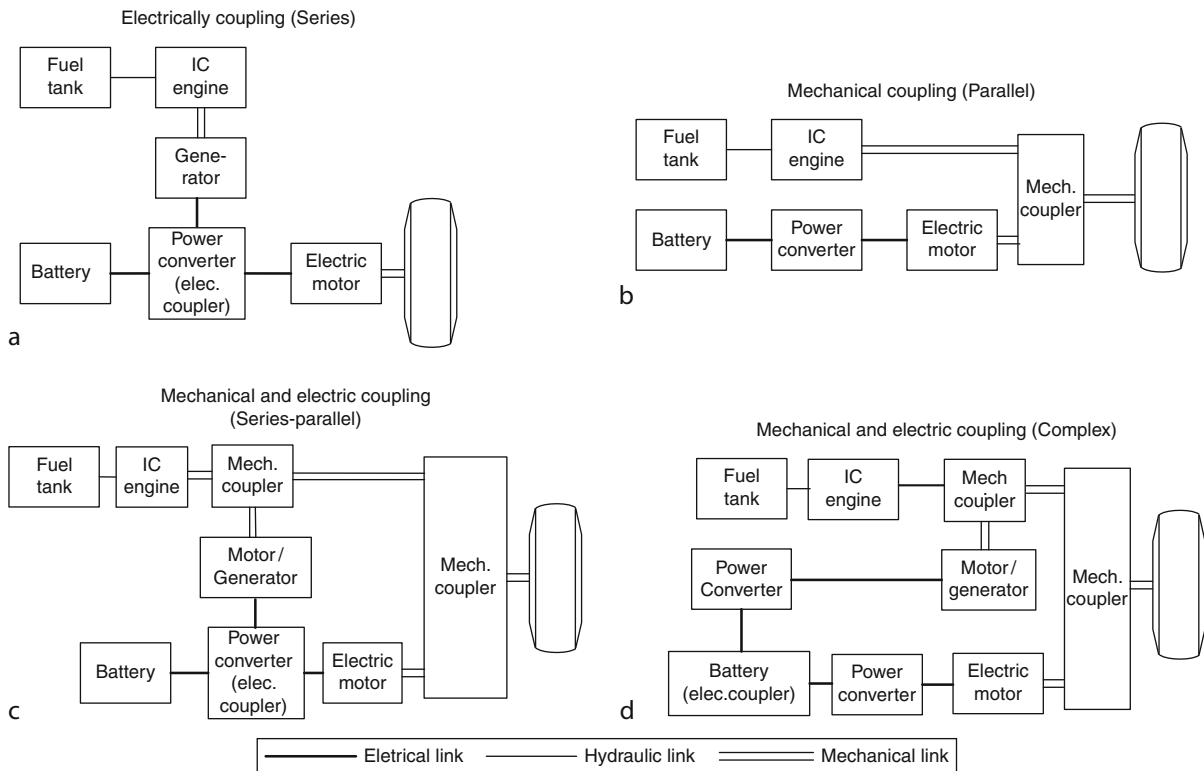
As discussed in a prior section, a hybrid vehicle has two power sources and one energy sink. Thus, there has to exist at least point at which two powers are merged together. At this point, the power forms have to be identical, mechanical or electrical. In this chapter, the terms of electric coupling and mechanical coupling are used rather than the traditional names. Figure 6 demonstrates the concepts of four configurations.

Figure 6a shows the configuration of an electric coupling, traditionally, a series configuration. Two powers are merged together at the power converter in the form of electrical. An internal combustion engine drives an electrical generator to supply electric power to the electronic converter, and at the same time, the battery pack also contributes its power to the power converter. The output power from the power converter flows to the electric motor to produce mechanical power and then drives vehicle wheels. The power shares are controlled by the power converter.

Figure 6b shows the configuration of mechanical coupling, traditionally, a parallel configuration. The internal combustion engine directly contributes its mechanical power to a mechanical coupler. At the same time, the battery pack supplies its power to the electric motor through a power converter which functions as the motor controller for controlling the motor torque. Two mechanical powers are merged together at the mechanical coupling and then delivered to drive wheels.

Figure 6c shows the configuration in which three power couplers are used, two mechanical and one electrical. There are plentiful available operation modes in the configuration that will be discussed in detail in following sections.

Figure 6d shows a configuration which is similar to Fig. 6c. The only difference is that the electric coupling function is moved from the power converter to the batteries and one more power converter is added between the motor/generator and the batteries [2].



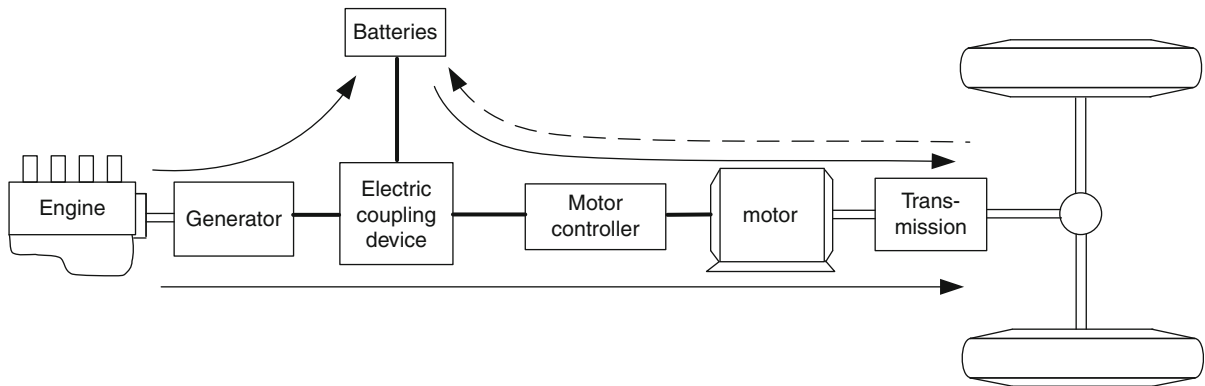
Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 6

Various configurations of hybrid vehicles [1]

Electric Coupling HEV (Series)

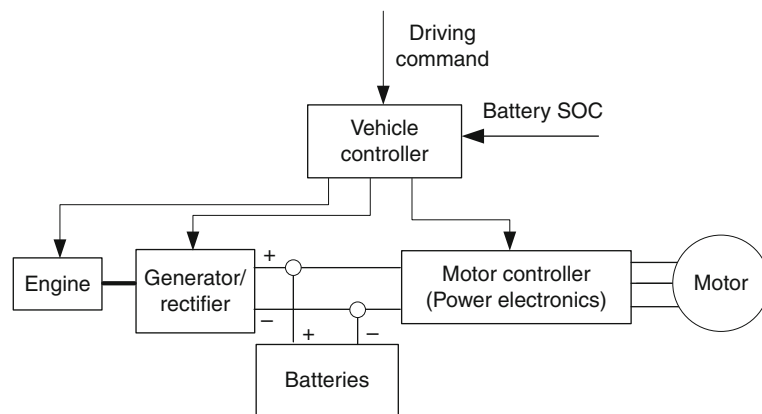
The electric coupling configuration (series) is schematically shown in Fig. 7, on which, the possible power flow routes are also shown. This drive train may have been developed originally from pure electric vehicles by adding a “range extender” engine/generator set. The engine/generator unit is used to contribute part of the electric energy for compensation of battery energy shortage. An electric power converter is used for energy management purpose based on the states of charge (SOC) of the battery pack. As shown in Fig. 7, there are several operation modes available:

1. Pure electric propelling mode, in which the engine/generator is shut down and only batteries are used to propel the vehicle.
2. Pure engine/generator traction, in which batteries are isolated from the power supply route and engine/generator power follows the load power. This operation mode is also used by diesel-electric locomotives
3. Hybrid propelling mode, in which both batteries and engine/generator supply power to the traction motor. The power shares between batteries and the engine/generator are controlled by coordinate controls of the power electronic converter and engine/generator.
4. Engine/generator power split mode, in which the power of engine/generator is separated by the power converter, driving wheels and charging batteries.
5. Regenerative braking mode, in which the electric motor is operated to function as electric braking to slow down the vehicle and get the braking energy back to the batteries. The engine/generator may be shut down meanwhile.
6. Battery charging mode, in which the engine/generator is operated only for charging batteries while vehicle stop.
7. Hybrid battery charging mode, in which both the engine/generator and the regenerative braking charge the batteries.



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 7

Electric coupling (series) hybrid drive train configuration and power flow scenario



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 8

Directly connected power sources and motor drive

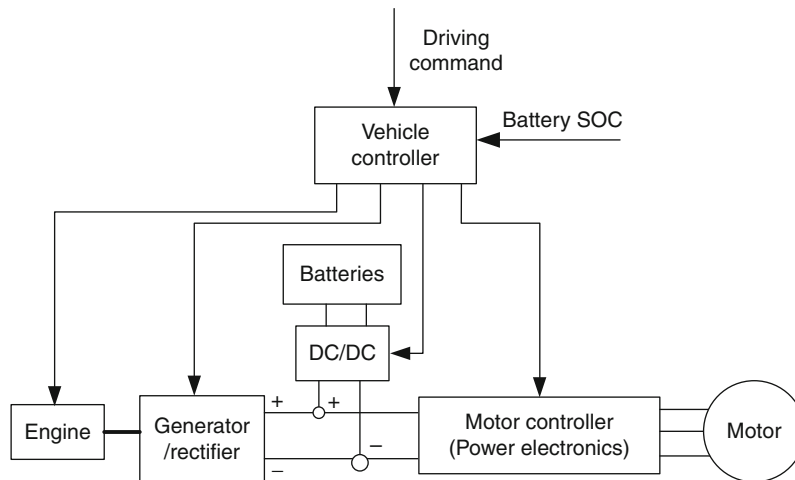
Electric Coupling Devices

The electric coupling device is the sole device at which two electrical powers are combined together. The operation modes discussed above can be carried out by coordinately controlling the electric coupling device and engine/generator. The simplest structure may be to directly connect the three terminals together as shown in Fig. 8.

The configuration in Fig. 8 has the features of identical voltage at the coupling point. The DC bus voltage is equal to the battery voltage. The power flows are only controlled by the voltage of the generator/rectifier, which depends on the operation speed of the generator and magnetic field. For battery charging, the engine/generator voltage has to be higher than the open

circuit voltage of the batteries. In regenerative braking, the bus voltage has to be higher than the battery voltage. However, the voltage generated by the motor in generating mode is proportional to the motor speed. Thus, the regenerative braking capability at low speeds may be rather limited.

A DC/DC converter can be added between the batteries and the DC bus to release the voltage constraint as shown Fig. 9. This design may significantly improve the performance of the drive train. It has the following advantages: (1) variations in the battery voltage do not affect the bus voltage, thus the engine/generator can be operated at a constant speed; (2) the battery energy can be fully used, which is very important for a plug-in hybrid drive train; (3) low



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 9
Configuration with a DC/DC converter

battery voltage and less battery cells can be used, which may lead to a small and light battery pack, low cost and high reliability; and (4) charging current of the batteries can be regulated during regenerative braking and charging from the engine/generator.

It is obvious that the DC/DC converter has to be bidirectional. It is considered to be the appropriate design with battery voltage lower than DC bus voltage. It boosts the battery voltage to match the DC bus voltage for propelling. On the other hand, it bucks the DC bus voltage to the level of the battery voltage for battery charging. In the situation of regenerative braking with the generated voltage by the motor higher than the battery voltage, the DC/DC operates in buck mode. However, at low speeds, the DC bus may be lower than the battery voltage. In this case, the DC/DC converter has to be operated in boost mode. The DC/DC operation mode is illustrated in Table 1 [2].

Figure 10 shows the topology of the four-quadrant DC/DC converter, operating in forward buck/boost and backward buck/boost by turning on or off the switches S_1 , S_2 , S_3 , and S_4 . The operation of the DC/DC converter is explained below.

Forward Bucking Forward operation is defined as the energy flow direction from batteries to DC bus. Forward bucking operation is that the DC/DC converter supplies current to the DC bus which has a voltage lower than battery voltage. In this operation,

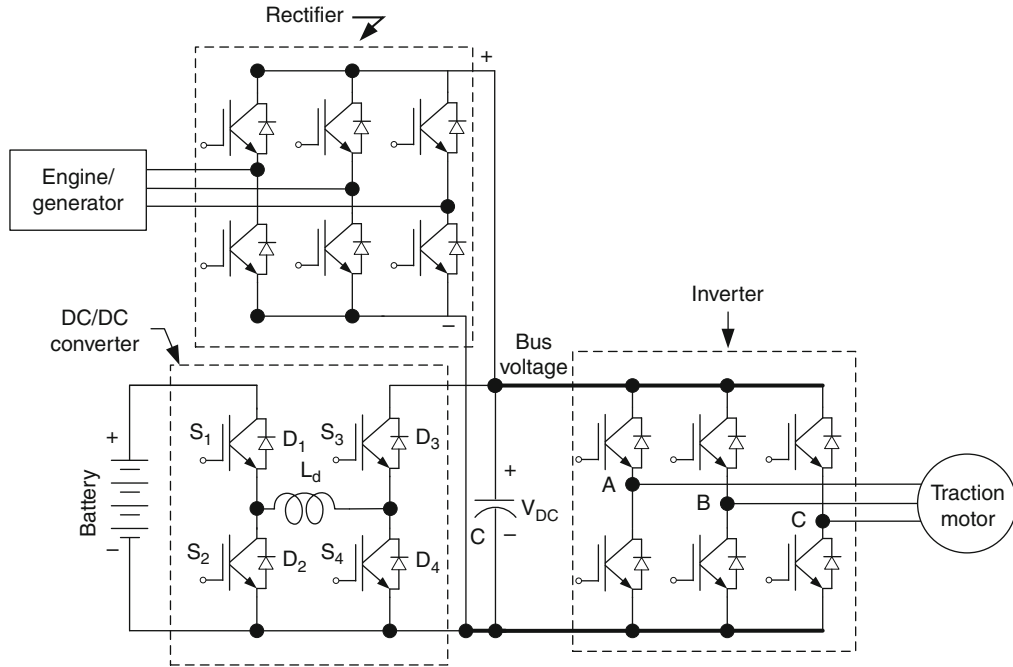
Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Table 1 Basic function a DC/DC converter [2]

	Energy flow	
	Battery discharging	Battery charging
Battery alone traction mode	Boost	–
Battery charging from engine/generator	–	Buck
Regenerative braking	–	Buck/boost

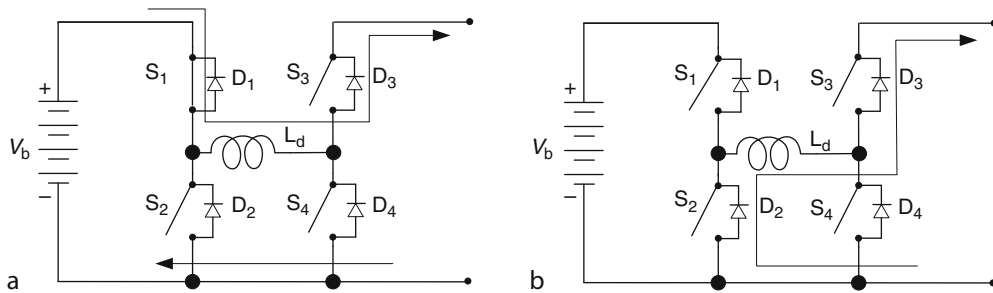
switches S_2 , S_3 , and S_4 are turned off as shown in Fig. 11. Switch S_1 is operated on and off periodically. During the on period of S_1 , the batteries supply current to the DC bus through S_1 , L_d , and D_3 as shown in Fig. 11a, and during the off period, the batteries are isolated from the DC bus, and the current continuously flows through D_2 , L_d , and D_3 . The bus voltage depends on the duty cycle δ , of S_1 , which is defined as the ratio of the on period, t_{on} , to whole period T ($T = t_{on} + t_{off}$). When $\delta = 0$, DC bus voltage is zero, and when $\delta = 1$, DC bus voltage equals to the battery voltage:

$$V_{bus} = \delta V_b. \quad (1)$$

The pulse width modulation (PWM) method is used to control the on and off operation of S_1 to obtain



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 10
Bidirectional DC/DC converter [2]



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 11
Forward bucking operation

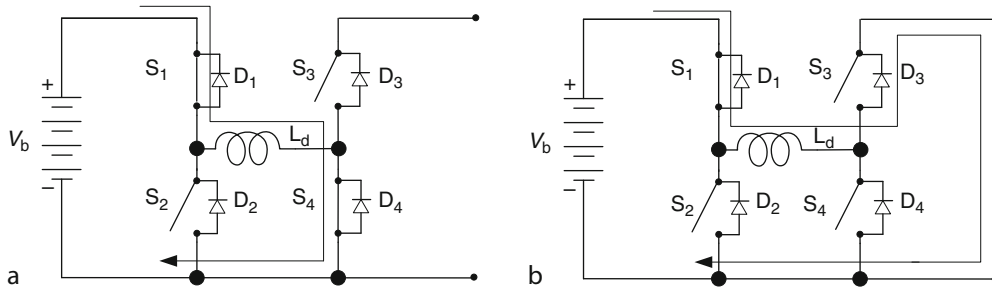
the desired DC bus voltage. This case may not be used with a design that the battery voltage is lower than the DC bus voltage.

Forward Boosting When battery voltage is lower than the DC bus voltage, the DC/DC converter is operated for boosting the battery voltage for supplying current to the DC bus. In this case, switch S_1 is on, S_2 and S_3 are off, and S_4 is operated on and off periodically as shown in Fig. 12. During the on period of S_4 ,

batteries charge the inductor L_d through S_1 and S_4 as shown in Fig. 12a. During the off period of S_4 , both batteries and the inductor L_d supply current to the DC bus, which has a higher voltage than batteries. The DC bus voltage depends on the duty cycle of S_4 :

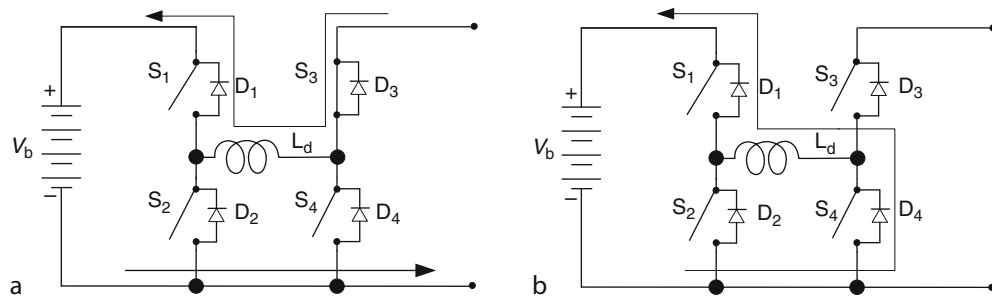
$$V_{\text{bus}} = \frac{V_b}{1 - \delta} \quad (2)$$

Backward Bucking Backward operation means the DC bus is supplying current to batteries from



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 12

Forward boosting operation



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 13

Backward bucking operation

regenerative braking or engine/generator. Bucking operation is to reduce the DC bus voltage to battery voltage level. In this operation, S_1 , S_2 , and S_4 are off and S_3 is operated on and off periodically. During the on period of S_3 , the DC bus supplies current to batteries through S_3 and D_1 as shown in Fig. 13a. During the off period of S_3 , the DC bus is isolated from the batteries. The batteries and DC bus have a voltage relation of

$$V_b = \delta V_{bus}. \quad (3)$$

This mode may be used in the situation of charging the batteries from engine/generator and regenerative braking at high vehicle speeds.

Backward Boosting This operation is to boost a low DC bus voltage to a high battery voltage. It may be used in regenerative braking at low vehicle speeds at which, the voltage generated by the motor is low. In this operation mode, S_1 and S_4 are off and S_3 is on and S_2 is operated on and off periodically. During the on period of S_2 , the DC bus charges the inductor L_d through S_3 and S_2 as shown in Fig. 14a. During the

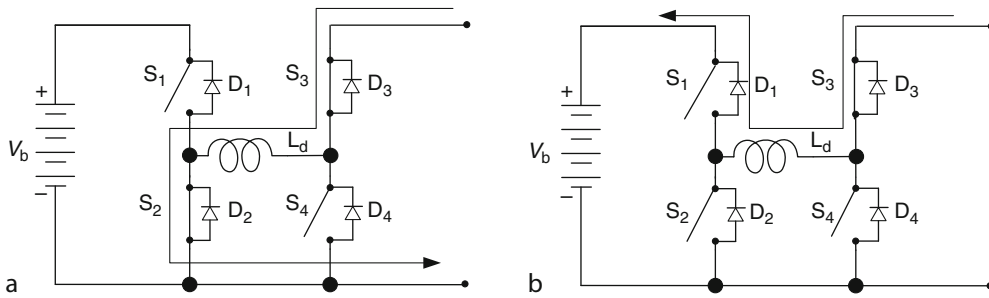
off period of S_2 , both the DC bus and inductor L_d supply energy to the batteries. The DC bus and batteries have the voltage relation:

$$V_b = \frac{V_{bus}}{1 - \delta}. \quad (4)$$

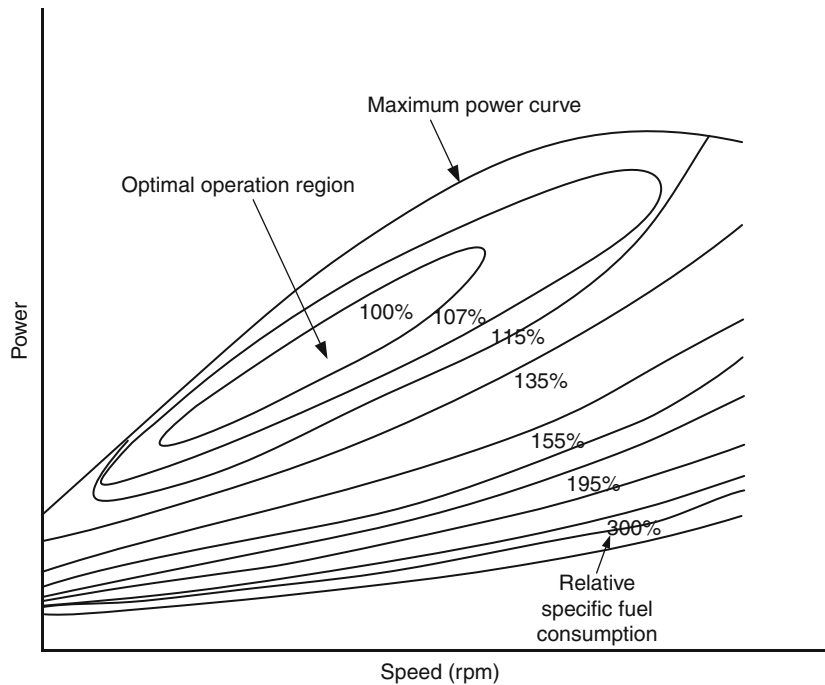
Advantages and Disadvantages of Electric Coupling (Series) Drive Train

Electric coupling (series) hybrid vehicles have the advantages of:

1. Internal combustion engine is completely mechanically decoupled from wheels. Thus the engine operation can be operated in its most efficient speeds range. On the other hand, since batteries supply supplementary power, engine power is also decoupled from load power. Therefore, the engine can be operated in its most efficiency power range as shown in Fig. 15.
2. Since the engine can be potentially operated in a very narrow speed and power region, it becomes much easier for further improvement of fuel



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 14
Backward boosting operation



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 15
Illustration of a typical engine fuel consumption map

- economy and emission characteristics by using some engine special design and control technologies in this narrow region than in the whole operation domain.
3. Completely mechanical decoupling of engine from driving wheels allows using engines that may not be suitable for direct vehicle propelling, such as gas turbine engine, sterling engine, etc.
4. A multi-gear transmission and torque converter may not be required, since the vehicle is solely driven by an electric motor, which has proper torque-speed profile for vehicle traction.
5. In-wheel motors may be used similar to a pure electric vehicle.
6. The control of the drive train may be simple, compared with other configurations.

However, electric coupling (series) drive train is also suffered from some disadvantages such as:

1. Two times energy form transformations from engine to wheels (mechanical to electrical and then to mechanical again) may cause significant energy losses.
2. Additional electric machine (generator) adds weight and cost.
3. Traction motor(s) have to be designed big enough to meet vehicle performance, such as maximum speed, acceleration, and gradeability.

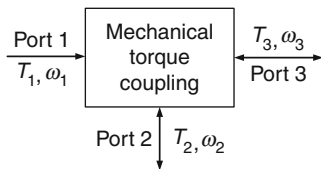
Mechanical Coupling Hybrid Drive Trains

In mechanical coupling hybrid vehicles, mechanical coupling devices are used to couple two mechanical powers together. Mechanical power composes of force and speed in linear motion, and torque and angular speed in rotary motion. Accordingly, a mechanical coupling device may add to torques together or add two angular speeds together, depending on its mechanical structure.

Torque Couplers A torque coupler is a three-port mechanical device which adds two input torques together and delivers them to output port as conceptually illustrated in Fig. 16. Port 1 is a unidirectional input port (e.g., engine port), port 2 is bidirectional input port (e.g., electric motor port), and port 3 is a bidirectional output port (e.g., wheel port).

Ignoring losses and inertia effect, the total input power is always equal to the output power. Supposing port 1 and port 2 are inputs and port 3 is output, one obtains:

$$T_3 \omega_3 = T_1 \omega_1 + T_2 \omega_2. \quad (5)$$



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 16

Concept of a mechanical coupling device

Inside the device, two torques, T_1 and T_2 , are added together in the following way:

$$T_3 = k_1 T_1 + k_2 T_2, \quad (6)$$

where k_1 and k_2 are structural parameters which are constants or variables, depending on the structure design of the device. T_1 and T_2 are independent of each other and then can be independently controlled. Solving Eqs. 5 and 6 gives the relationship among the speeds:

$$\omega_3 = \frac{\omega_1}{k_1} = \frac{\omega_2}{k_2} \quad (7)$$

Obviously, three speeds are linked together by k_1 and k_2 , and have a fixed relationship with each other. When a torque coupler is used in a hybrid vehicle, the engine torque can be decoupled from the wheel torque and motor torque is used for compensation. However, the engine speed varies with the vehicle speed. This feature provides the engine with an opportunity of always operating at its operation line at any vehicle load as shown in Fig. 17. Figure 18 shows the commonly used torque couplers.

Speed Couplers Similar to the torque coupler (also refer to Fig. 16), a mechanical speed coupler can be described as:

$$\omega_3 = k_1 \omega_1 + k_2 \omega_2 \quad (8)$$

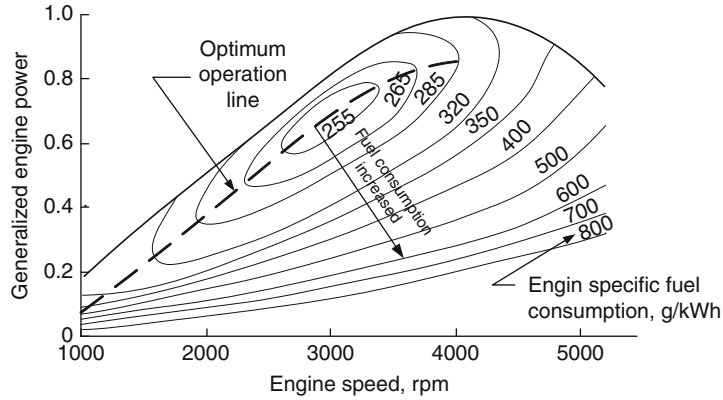
where k_1 and k_2 are the constants or variables associated with the device structure design. Among these three speeds, two of them are independent and then can be controlled independently. Combining with Eq. 5, the torques are linked together by:

$$T_3 = \frac{T_1}{k_1} = \frac{T_2}{k_2} \quad (9)$$

A typical speed coupler is the planetary gear unit as shown in Fig. 19. It consists of a sun gear 1, a ring gear 2, and a yoke 3 which is hinged to the centers of several planetary gears (usually three or four gears for force balance). The relationship among ω_1 , ω_2 , and ω_3 is

$$\omega_3 = \frac{1}{1 + g_r} \omega_1 + \frac{g_r}{1 + g_r} \omega_2 \quad (10)$$

where g_r is the gear ratio of ring gear to sun gear ($R_2/R_1 = z_2/z_1$, where z_2 and z_1 are the tooth number



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 17
Engine fuel consumption map and its optimal operation line

of ring gear 2 and sun gear 1). Obviously, constants k_1 and k_2 in Eqs. 8 and 9 are $k_1 = 1/(1+g_r)$ and $k_2 = g_r/(1+g_r)$. The directions of the torques as shown in Fig. 19 (counterclockwise for sun and ring gears and clockwise for yoke) are defined as positive. The torque relationship can be expressed by Eq. 9. It can be seen from Fig. 19 that the torques on sun gear and ring gear always have the same direction and the torque on the yoke always has an opposite direction.

Another interesting speed coupling device is an electric motor with a floating stator [2], in which the stator, generally fixed to a stationary frame in conventional motor, is released and forms a double rotor motor as shown in Fig. 20. In this device, port 2 is an electric port, at which electric power is converted into mechanical power at the air gap between the inner and outer rotor to generate a relative speed ω_2 . Obviously,

$$\omega_3 = \omega_1 + \omega_2, \quad (11)$$

and

$$T_3 = T_1 = T_2. \quad (12)$$

Here, $k_1 = 1$ and $k_2 = 1$ in Eqs. 8 and 9.

With a speed coupler, the engine speed is decoupled from the vehicle speed, thus the engine can always be operated within its optimal speed range.

Selective uses of electric coupling, mechanical torque, and speed couplings can create many hybrid vehicle configurations as discussed in following sections.

Hybrid Drive Train with One Torque Coupler

Hybrid drive train with one torque coupler is the simplest configuration which is traditionally referred to as parallel configuration. The torque coupler can be designed with different k_1 and k_2 values as indicated in Eqs. 6 and 7 and Fig. 16. A two-shaft configuration with gear box as its coupler and two transmissions is shown in Fig. 21.

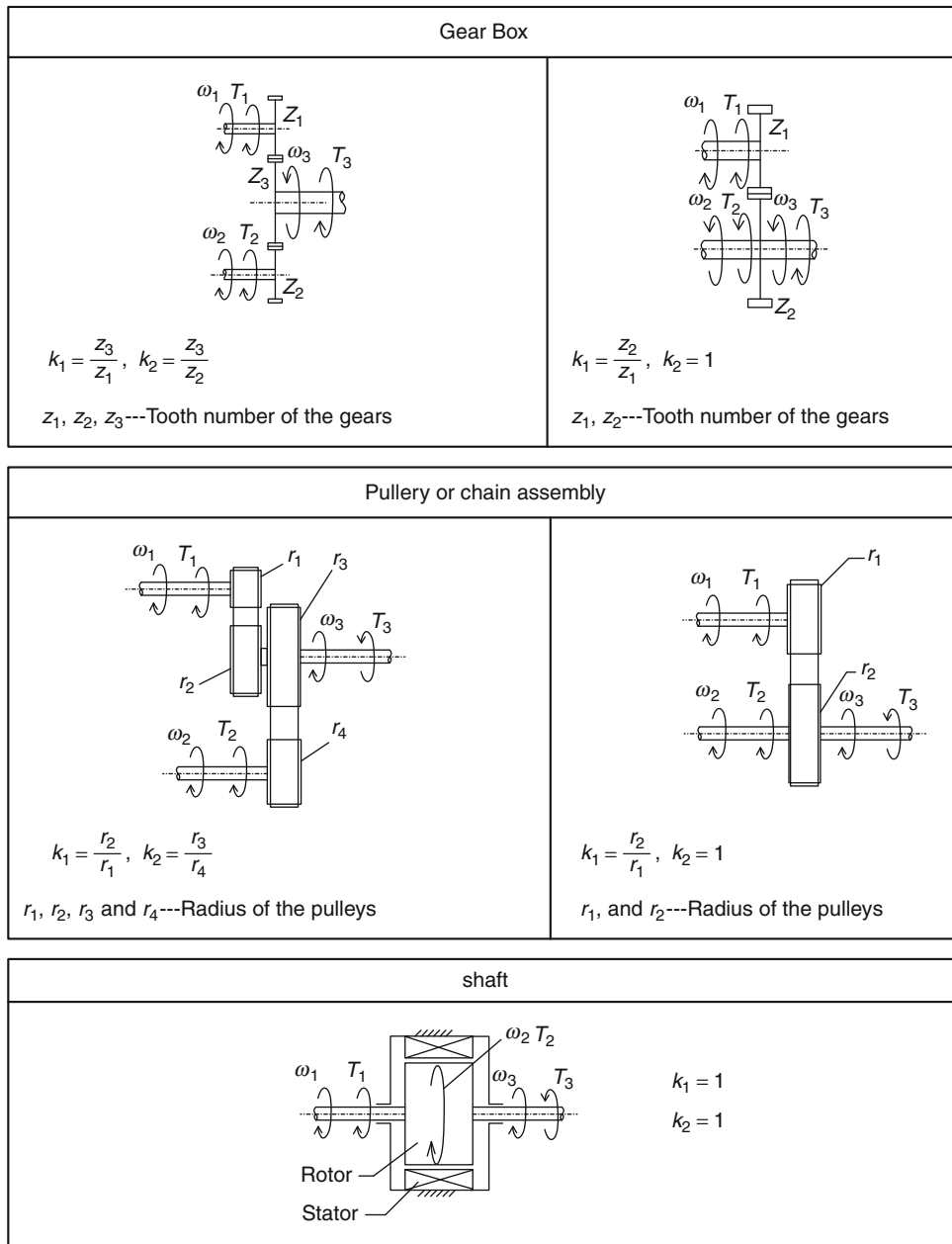
In this configuration, the torque coupler can be designed with any parameter values of k_1 and k_2 . In front of the torque coupler, two transmissions can be used, which may be multi-gear, single gear, or continuous varying transmission (CVT), depending on the engine and motor torque-speed characteristics and vehicle performance requirements. Actually, two transmissions and the torque coupler together can be taken as one assembly, which has the inputs of engine and motor and the output to final drive. The torque coupling assembly still has the same characteristic of the single torque coupler explained by Eqs. 6 and 7, except gear ratios of two transmissions being added, which can be expressed as:

$$T_o = k_1 i_e T_e + k_2 i_m T_m. \quad (13)$$

and

$$\omega_o = \frac{\omega_e}{k_1 i_e} = \frac{\omega_m}{k_2 i_m}, \quad (14)$$

where ω_e , T_e , ω_m , and T_m are the angular speeds and torques of the engine and motor, i_e and i_m are the



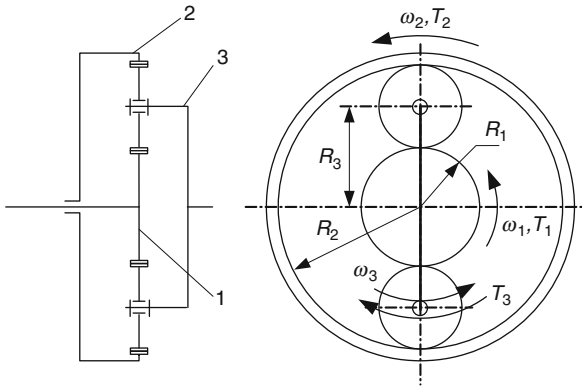
Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 18

Commonly used torque coupling devices [2]

transmission gear ratios of the engine and motor transmissions, respectively.

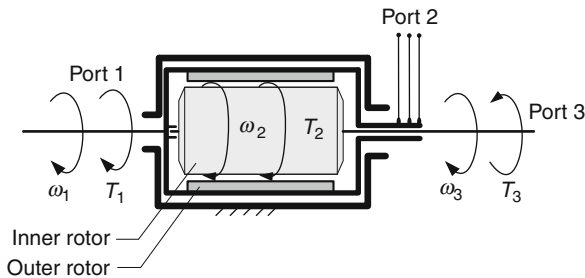
Obviously, two multi-gear transmissions produce many tractive effort profiles [2]. For example, two three-gear transmissions produce nine tractive effort profiles (3×3). High vehicle acceleration and

gradeability performance can be obtained. Further, high overall efficiency of the drive train can also be obtained by selecting proper gears for operating the engine and motor within their best region. This design also provides opportunity for using the engine and motor which may have less favorable characteristics.



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 19

Planetary gear unit used as a speed coupler [2]



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 20

Electric motor with double rotors as a speed coupler [2]

However, this design makes it complicated for the control system for selecting proper gears.

Instead, a single gear transmission for the motor may be used, since electric motor generally has good torque-speed characteristics. The three-gear transmission for engine is used to modify the engine torque-speed profile and provides opportunity for engine operating closer to its high-efficiency region.

Another possible design is to use single gear transmission for engine and multi-gear transmission for engine. This design is considered to be an unfavorable design because it does not take advantage of the two power plants.

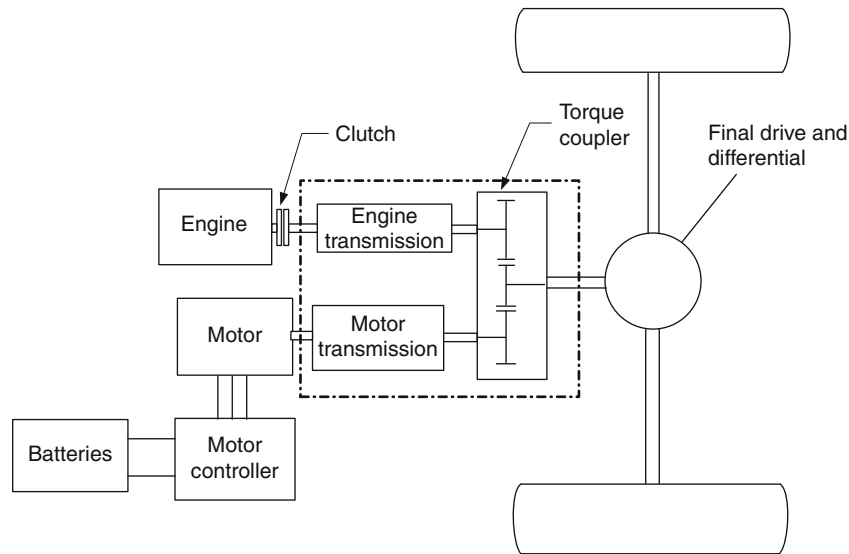
The simplest design is to use two single gear transmissions. When the engine and motor are designed to meet gradeability and acceleration performance, it is considered to be a good design.

Another two-shaft torque coupling hybrid configuration is to place a transmission behind the torque coupler as shown in Fig. 22 [1], in which, the engine and motor torque-speed profiles and the tractive effort are also illustrated. Obviously, the transmission modifies the engine and motor torques with the same scale. The gear ratio of the torque coupler (k_1 and k_2) is designed to matches the speeds of the engine and motor. This design may be a proper selection for the one with larger engine and small motor, such as mild hybrid drive train.

A simple shaft may be used as the torque coupler, which constitutes the configuration, traditionally called one-shaft parallel drive train. A transmission may be placed in front of the torque coupler as shown in Fig. 23. In this configuration, the rotor of the electric motor functions as the torque coupler, in which one end is connected to the engine and the other end to the transmission. Obviously, here, $k_1 = 1$ and $k_2 = 1$. This configuration has the similar operation characteristics. Since the motor directly drives the final drive, the motor speed is proportional to the vehicle speed, and the final drive gear ratio should be designed to match the motor speed and vehicle speed.

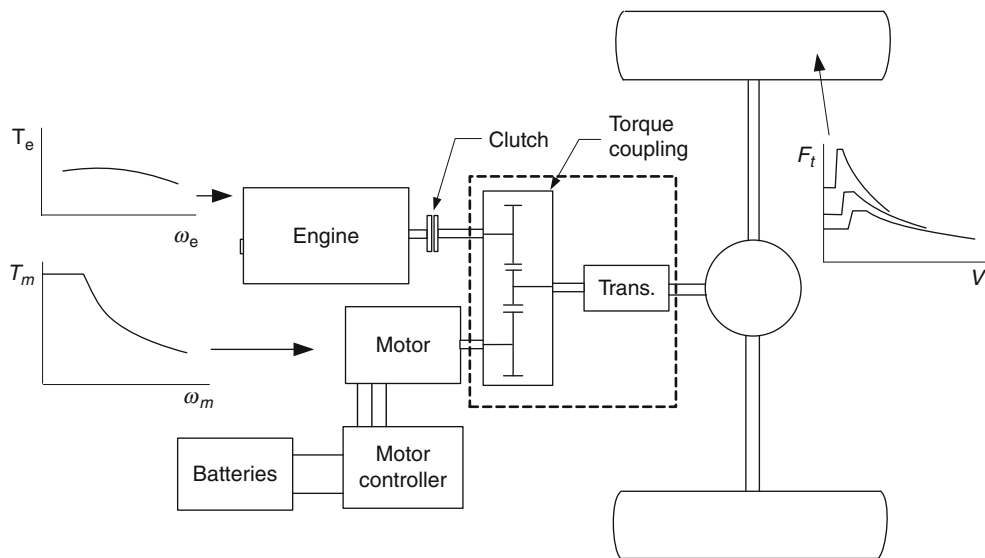
An alternative configuration is to place the transmission behind the motor as shown in Fig. 24. In this configuration, torques of both engine and motor are enhanced by the transmission. However, the engine and motor are directly coupled together; hence, they have to have the same operation speed range. The motor can serve as the engine start when a clutch is placed between the motor and transmission. When this configuration is used in a mild hybrid drive train, where motor alone operation mode is used, the clutch can be placed between the motor and transmission. The motor functions as an engine torque booster, engine start, generator, and regenerative braking.

Actually, the vehicle chassis and road can also used as a torque coupler. Figure 25 shows a hybrid drive train in which the engine and motor drive the front and rear wheels, respectively. The tractive efforts produced by the engine and motor are added together by the vehicle chassis and road. This configuration offers some advantages of keeping the original engine and transmission driveline and providing the vehicle with a four-wheel-drive capability.



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 21

General torque coupling hybrid drive train [2]



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 22

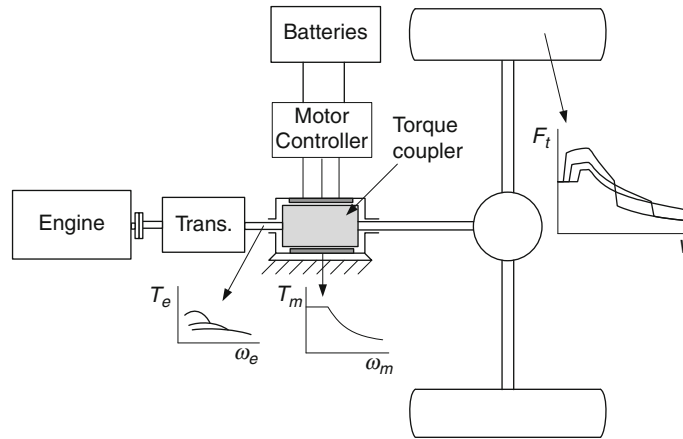
Two-shaft configuration with a transmission behind the torque coupler [2]

Hybrid Drive Trains with Torque and Speed Couplings

As discussed in prior sections, the torque coupler can decouple the engine torque from the wheel's torque demand, and speed coupler can decouple the engine speed from the wheel speed. Thus, it is possible to

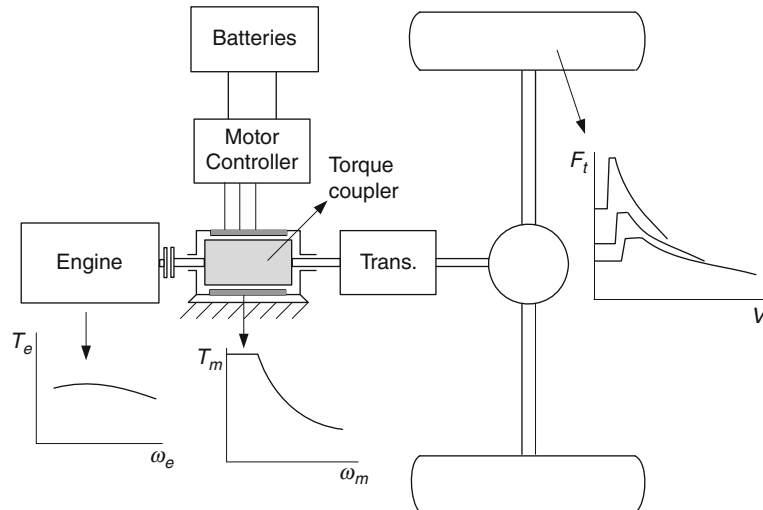
decouple the engine torque and speed from the wheels with two couplers. This concept is illustrated in Fig. 26.

A detailed drive train structure is shown in Fig. 27 [1]. The drive train uses a planetary gear unit as its speed coupler. Engine is connected to its yoke and



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 23

One-shaft configuration with a transmission in front of the torque coupler [1]



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 24

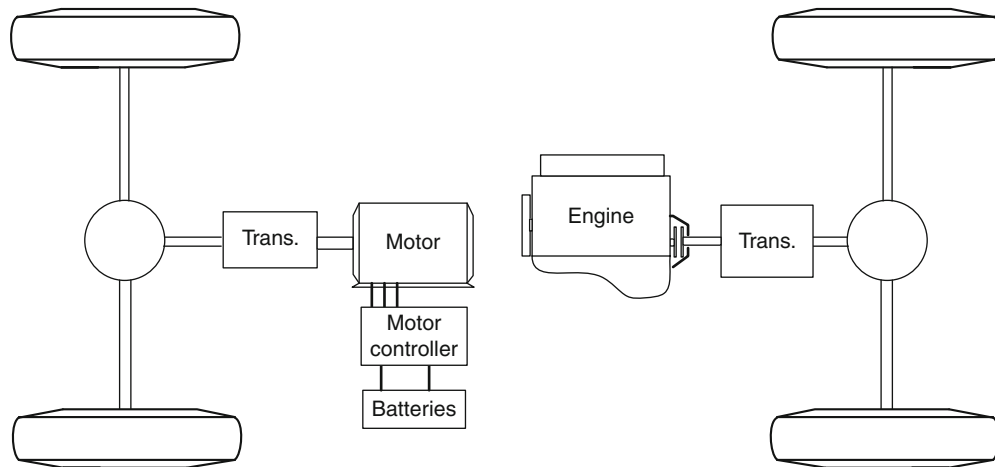
One-shaft configuration with a transmission behind the torque coupler [2]

a motor/generator to its sun gear. The ring gear as the output port is connected to a torque coupler which is a fixed shaft gear set. Another two ports of the torque coupler are connected to a traction motor and wheels.

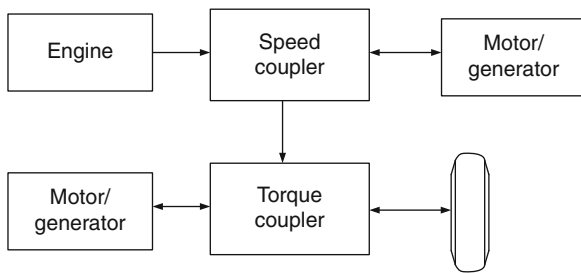
As indicated in Eqs. 8 and 9, Fig. 19, the torque of the motor/generator always has an opposite direction as that of the engine and can rotate in two directions. When the rotating speed is in the same direction as its torque, it is a motor, otherwise, a generator. When it operates in the generating mode, it absorbs part of the engine power.

Due to this reason, this hybrid drive train gets another name of engine power split hybrid drive train.

The torque of the motor/generator is always proportional to the engine torque. Thus, an ideal operation characteristic is a constant torque at the whole positive and negative speed domain. Also, the three controllable components (the engine, motor/generator, and traction motor) provide very plentiful operation modes, and hence, greater opportunity for high operating efficiency. However, very complex control is required [2].



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 25
Front- and rear-wheel-driven hybrid vehicle



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 26
Concept of torque and speed coupling drive train

Architectures of Fuel Cell Vehicles

Fuel Cell System

A fuel cell is a chemical device that converts the chemical energy into electric energy by means of redox reaction with the operation principle exactly the same as in a chemical battery. In a battery cell, the reactive chemicals are pre-installed in the cell. When the active chemicals are used up, the cell is dead and has to be recovered by charging it from an outside electrical source. Instead, a fuel cell is fed with its reactive chemicals continuously from outside of the cell. It is more like an internal combustion engine which is fed with fuel and produces mechanical energy. However, the energy conversion efficiency in a fuel cell is higher than internal combustion engine due to its free from the Carnot cycle limitation.

Compared to chemical batteries, it has the advantages of quick refueling and much more energy in an outside energy storage, similar to a fuel tank.

A fuel cell has a very similar structure of chemical battery cell, as shown in Fig. 28. Fuel and the oxidizing agency are continuously and separately fed into the anode and cathode electrodes, where they are ionized. Electrolyte is used to conduct positive ions from the anode to the cathode, and at the same time, electrons are conducted from the anode to the cathode through an electrical load.

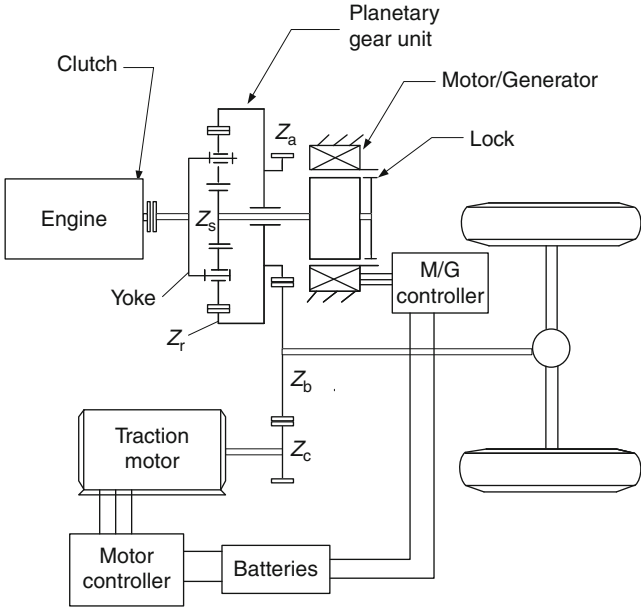
A single fuel cell has very low voltage. In practice, a fuel cell stack may include hundreds of single cells serially connected together. A fuel stack needs auxiliaries to support its operation. The auxiliaries mainly include air circulating pump, coolant circulating pump, ventilation fan, fuel supply pump, and electric control devices. Part of the power developed from the fuel cell stack is used to support the operation of the auxiliaries as shown in Figs. 29 and 30.

Fuel Cell Technologies

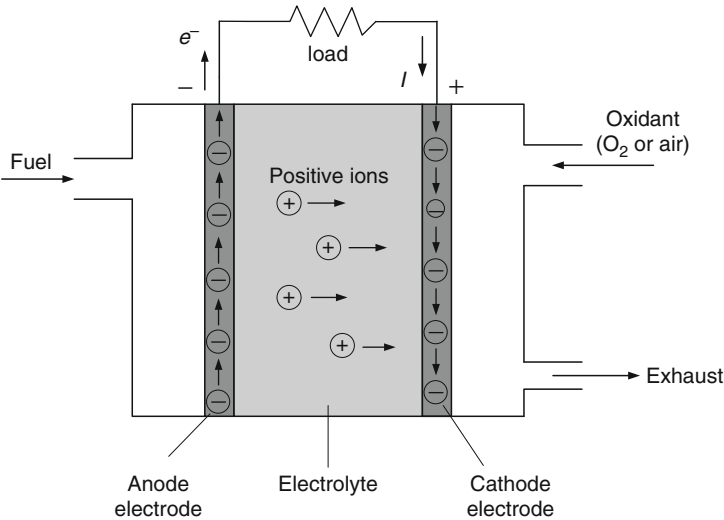
There are several types of fuel cells, depending on their electrolytes. Table 2 lists the operation temperature and electrolytes.

Proton Exchange Membrane Fuel Cells (PEMFCs)

PEMFCs work with hydrogen as fuel and oxygen as



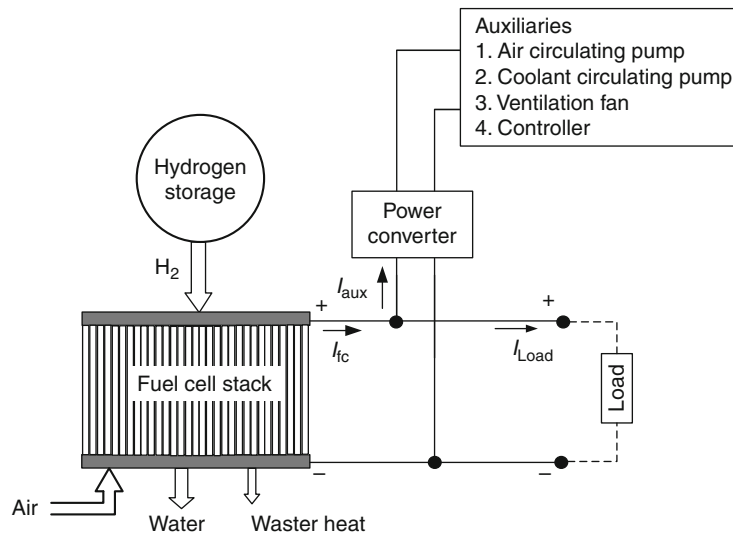
Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 27
Hybrid drive train with speed and torque couplers



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 28
Illustration of fuel cell operation

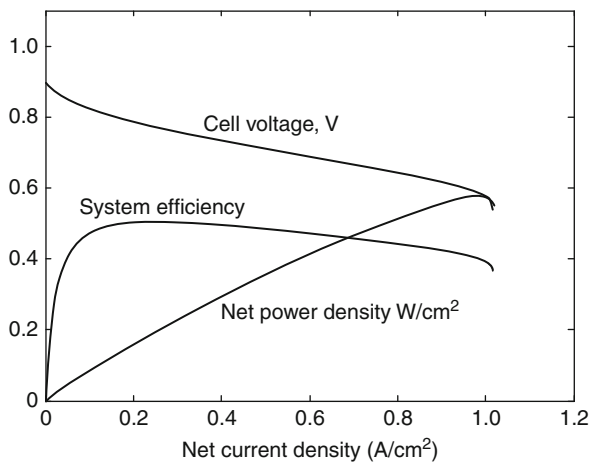
oxidant. Hydrogen is supplied to the anode where it is oxidized, losing electrons in the process. Positive hydrogen ions (protons) migrate across the electrolyte through the membrane to the cathode, at the same

time, the electrons travel round the external circuit to the cathode. The oxygen is supplied to the cathode where it is reduced, picking up the electrons and the ions from the hydrogen to form water.



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 29

A hydrogen–air fuel cell system



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 30

Cell voltage, system efficiency, and net power density varying along net current density of a hydrogen–oxygen fuel cell [2]

The electrodes of PEMFCs are made of porous carbon which allows the active gases to pass through. The electrode surfaces are coated with platinum as a catalyst which increases the rate of oxidation at the anode and the rate of reduction at the cathode. The platinum catalysts allow the chemical reaction to take place at a lower temperature. Alternatively to avoid

Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Table 2 Operating data of various fuel cell systems [3]

Cell system	Operating temperature, °C	Electrolyte
Proton exchange fuel cells (PEMFCs)	60–100	Solid
Alkaline fuel cells (AFCs)	80–230	Liquid
Phosphoric acid fuel cells (PAFCs)	60–200	Liquid
Molten carbonate fuel cells (MCFCs)	600–1000	Liquid
Solid oxide fuel cells (SOFCs)	1000–1200	Solid
Direct methanol fuel cells (DMFCs)	100	Solid

the cost of expensive catalysts, some fuel cells are designed to work at elevated temperatures.

The platinum catalyst used in PEMFCs is very expensive and extremely sensitive to poisoning by even small amounts of carbon monoxide. Thus a high purity of hydrogen gas is necessary.

Electrolyte is a thin, fragile sheet of acidic, solid organic polymer about 50 μm thick which permits the

passage of hydrogen ions but is impermeable to electrons. Acidic compounds are fluids with free hydrogen ions and these are the fuel cell's charge carriers [4].

The first PEMFCs were developed in the 1960s for the need of a US-manned space program. It is nowadays the most investigated fuel cell technology for automotive applications. It operates at 60–100°C and can offer a power density of 0.35–0.6 W/cm². PEMFCs offer some definite advantages for electric and hybrid vehicle applications that include low temperature operation, high power density among all the available types of fuel cells, solid electrolyte that does not change, movement and vaporization from the cell, and most importantly, potentially high efficiency and being free of toxins and greenhouse gas emissions. The major problems are their requirements of expensive noble metal catalyst and membrane, and a need of high purity of hydrogen as its fuel.

Alkaline Fuel Cells (AFCs) Alkaline fuel cells use aqueous electrolytes of potassium hydroxide (KOH) to conduct ions between electrodes. Hydrogen and oxygen are used as the fuel and oxidant. The ion carried by the alkaline electrolyte is a hydroxide ion (OH[−]). The half reactions are:



AFCs are capable of operating over a wide range of temperature (80–230°C) and pressure (2.2–45 atm). AFCs are capable of achieving high operating efficiency because of the fast kinetics permitted by the hydroxide electrolyte. The fast kinetics in AFCs allows using silver or nickel as catalysts instead of expensive platinum.

The performance of AFCs can be further improved by circulating the electrolyte, which offers some advantages that include easy thermal management, more homogenous electrolyte concentration, easy replacement of electrolyte, and possible removal of the electrolyte when it is turned off. Of course, circulation of electrolyte needs additional equipments, such as circulation pump, heat exchanger, and evaporator. Further, it increases the risk of leakage (potassium hydroxide is highly corrosive and has a natural tendency to leak even through the tightest seals).

AFCs were some of the earliest practical cells and were used in the Apollo space program, generating drinking water as well as electric power. Although they are inexpensive compared with PEM cells, operating efficiencies of 60% are possible. Unfortunately they have a low power output and the catalyst is prone to poisoning from carbon dioxide in the atmosphere.

Phosphoric Acid Fuel Cells (PAFCs) PAFCs use acidic electrolyte to conduct hydrogen ions and the anode and cathode reactions are the same as in PEMFCs. PAFCs run at high temperature of around 200°C, delivering high power of a MegaWatt or more but with relatively low efficiency of around 35%. The poor conversion efficiency is the high heat generation in the fuel cell stack. Because of the high working temperature the efficiency losses can be mitigated by using the waste heat in combined heat and power (CHP) applications.

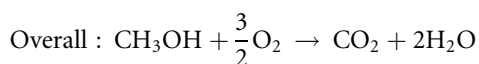
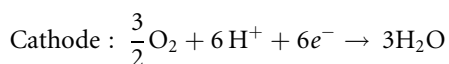
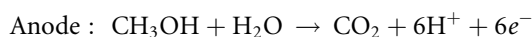
The advantage of PAFCs is the cheap electrolyte and disadvantages are the need of expensive catalyst (platinum), corrosion by acidic electrolyte, CO poisoning, and low efficiency.

Molten Carbonate Fuel Cells (MCFCs) MCFCs run at high temperatures of 650–1,000°C. They use molten carbonate salt to conduct ions, usually, lithium-potassium carbonate or lithium-sodium carbonate. Their unique chemistry needs carbon dioxide from the air as part of the process. Efficiencies achieved are 45% or more and power outputs of over 1 MW are typical in grid supply applications. Because of their high working temperature, they can operate directly with hydrocarbon gases which are reformed within the cell and do not need a separate hydrogen supply. The high temperature also means that less expensive catalysts are needed, but the molten electrolyte imposes special requirements on containment and anticorrosion measures.

Solid Oxide Fuel Cells (SOFCs) Solid oxide fuel cells also operate in the same or higher temperatures as the molten carbonate cells with the same fuel and catalyst advantages. The ceramic electrolyte which can run as hot as 800°C has the advantage that the electrolyte stays solid. They can deliver powers of several Megawatts but at a lower efficiency of around 35%.

Direct Methanol Fuel Cells (DMFCs) Direct methanol fuel cells are similar to PEM cells. Instead of pure hydrogen, they directly use liquid Methanol as fuel. It is sure that there are definite motivations for using DMFCs to vehicles. Methanol is a liquid fuel at room temperature that can be easily stored, distributed, and marketed through current infrastructures. Further, methanol is the simplest organic fuel that can be economically and efficiently produced on a large scale from fossil fuel, such as coal and natural gas. Further it can be also produced from biomass products.

Anode and cathode electrode of DMFC also need platinum or platinum alloys as catalyst. The electrolyte can be a trifluoromethane sulfonia acid or proton exchange membrane (PEM). The chemical reactions in a DMFC are:



Like PEM fuel cells, DMFCs work at low temperatures in the range of about 50–120°C. They have a relatively low efficiency and power density. Output power using current technology is limited to about 1.5 kW which is enough to power most consumer goods but not sufficient for automotive applications.

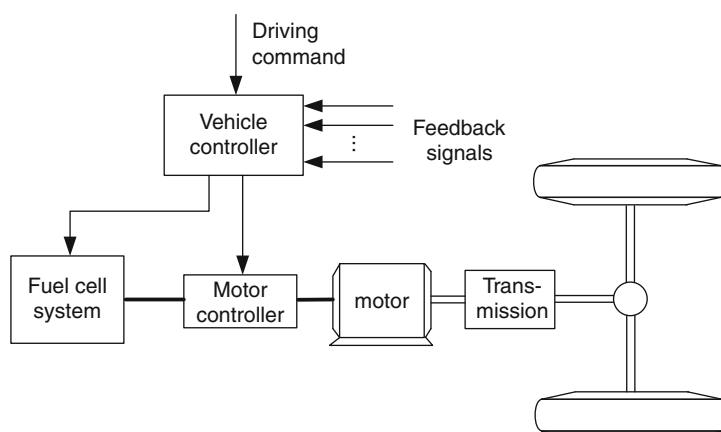
Nevertheless the ability to use liquid fuel coupled with the elimination of the reformer make these fuel cells very attractive.

Fuel Cell Vehicles

Fuel cell-powered vehicles potentially have the advantages of no toxic and greenhouse emissions and high operation efficiency. Compared to internal combustion engine, fuel cells have a prominent advantage of high operation efficiency at low power output which is coincident to a vehicle's normal driving situation. This means that fuel cells in a fuel cell-powered vehicle will naturally operate in their high-efficiency region in normal driving conditions.

Fuel cell-powered vehicles can be either fuel cell alone or hybridized with batteries and/or ultra-capacitors. Fuel-cell-alone-powered vehicles have a very simple configuration, similar to pure electric vehicles by replacing the battery pack with a fuel cell system as shown in Fig. 31.

A fuel-cell-alone-powered vehicle is a single power source vehicle. The fuel cell power follows the load profile. The power rating of the fuel cell system is designed to meet the vehicle performance requirement. A fuel cell system management is required to manage the fuel cell operation, such as air supply, fuel supply, and cooling of the fuel cells for operating in their safe range.



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 31
Configuration of fuel-cell-alone-powered vehicle drive train

Fuel cell is a unidirectional energy converter. It cannot operate in opposite directions to absorb energy from an electric energy source. Thus, a fuel-cell-alone-powered vehicle is not capable of recovering braking energy.

Fuel Cell–Powered Hybrid Vehicles

Fuel cell system, at today's technology, has low power density. For satisfying vehicle performance, a bulky and heavy fuel cell system is needed, which in turn, harms the vehicle performance and causes high energy losses. In addition, lack of regenerative braking capability is a major shortcoming of this drive train.

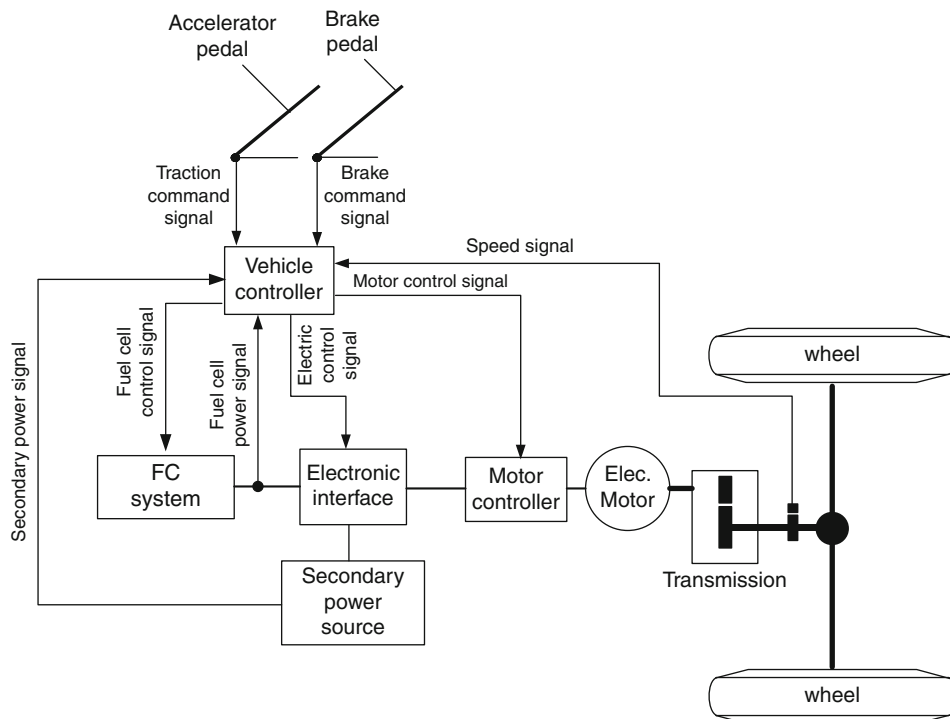
Hybridization of fuel cell system with a secondary power source, such as a battery pack and/or ultracapacitor pack is an effective technology to overcome these shortcomings. A fuel cell hybrid vehicle drive train is shown in Fig. 32.

Fuel cell hybrid vehicle drive train consists mainly of a fuel cell system as primary power source, a secondary power source that may be a battery pack

and/or ultracapacitor pack. The powers of the fuel cells and secondary power source are coupled together at an electronic interface which is used to control the power flows. A vehicle controller is used to control the drive train based on the driver's command and system's feedback signals.

There are several operation modes that can be selectively used which are:

1. Fuel cell alone propelling: in which the electronic interface isolates the secondary power source from the power line and fuel cell pack alone supplies power to the electric motor.
2. Secondary power source alone propelling: in which the fuel cell pack is isolated from the power line and the secondary power source alone supplies power to the traction motor.
3. Hybrid propelling: in which both the fuel cell system and the secondary power source provide powers to traction motor simultaneously.
4. Fuel cell power split mode: in which fuel cell power is split into two parts by the electronic interface.



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 32
Configuration of a fuel cell–powered hybrid drive train

One part is used to power the traction motor and the other charges the secondary power source.

5. Regenerative braking: in which the traction motor functions as a generator to recover the braking energy and stores it in the secondary power source.

By electively using the operation modes, the fuel cells can be operated in their most efficient region as shown in Fig. 33.

The electronic interface plays a critical role in the drive train control. It should be designed with three ports and the ports to the secondary power source and electric motor are bidirectional. The configuration and operation principle are very similar as that used in engine/generator-based electric coupling (series) hybrid drive train.

Future Directions

The electric vehicle has a relatively simple power train structure. The most important components are the motor drive and the battery system. An advanced motor drive with high power density, high efficiency, and long extended speed range is the key technology. The success of an electric vehicle is almost determined by the performance of the battery system. The basic requirements for electric vehicle batteries are the high energy capacity for extended range, high power for

vehicle performance, high safety, and low cost. At present, lithium-ion batteries are considered to be the most promising candidates. For hybrid vehicles, advanced primary power sources are very important that may include advantage engine technologies and alternative fuel engines. Advanced control strategies and real-time control algorithms are also very important. The key technologies for fuel cell vehicles include enhancement of fuel cell power density, reduction of cost, and, more importantly, hydrogen production and onboard storage.

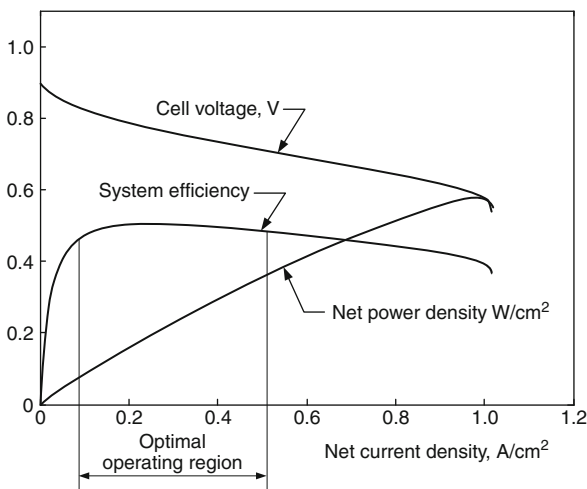
Bibliography

Primary Literature

1. Chan CC, Chau KT (2001) Modern electric and vehicle technology. Oxford University Press, New York
2. Ehsani M, Gao Y, Emadi A (2010) Modern electric, hybrid electric and fuel cell vehicles—Fundamentals, theory and design, CRC, FL, USA
3. Berlowitz PJ, Darnell CP (2001) Fuel choice for fuel cell powered vehicles. Society of automotive engineers (SAE), Journal, Paper No. 2001-02-0003
4. http://www.mpoweruk.com/hydrogen_fuel.htm. Accessed Oct 2010

Books and Reviews

- Miller JM (2004) Propulsion systems for hybrid vehicles. The Institution of Electrical Engineers, London
- Messlerle HK (1969) Energy conversion statics. Academic, New York
- Larminie J, Dicks A (2000) Fuel cell system explained. Wiley, New York
- Thomas CE, James BD, Lomax FD Jr, Kuhn IK Jr (2002) Societal impact of fuel options for fuel cell vehicles. SAE Journal, Paper No. 982496



Electric, Hybrid Electric and Fuel Cell Vehicles, Architectures of. Figure 33

Optimal operation region of a fuel cell system

Electric, Hybrid, and Fuel Cell Vehicles, Introduction

MEHRDAD EHSANI

Department of Electrical & Computer Engineering,
Texas A&M University, College Station, TX, USA

The success of internal combustion engine vehicles over the past century is one of the greatest technological achievements of mankind with a profound impact

on virtually every aspect of human life on the planet. However, the very widespread success of this technology has produced associated problems which can be the seeds of its own long-term demise. Solution of these problems will guarantee the long-term survival of personal and mass transit vehicles which are so essential to the economy and welfare of human beings. The proper combination of these solutions can bring about what is termed sustainable vehicles.

This section of the encyclopedia describes the important sustainable vehicle technical topics and the present state of the art on each topic. The following is a brief review of the entries in this section.

► [Sustainable Transportation](#) describes the technical and environmental problems that have to be addressed to achieve sustainable vehicles for the future. Issues of fuel, environment, and vehicle technologies are introduced.

► [Electric, Hybrid Electric and Fuel cell Vehicles, Architectures of](#) describes the basic configuration of the drivetrain of electric vehicles and hybrid electric vehicles. The various types of hybrid electric vehicles' drivetrains are presented. The basic operation of each of these systems and their components are described.

► [Vehicle Dynamics and Performance](#) describes vehicle drivability, fuel economy, braking performance, handling characteristics, noise, vibration, and harshness. Vehicle performance discussed in this entry is restricted to propelling and braking performance in terms of vehicle speed, acceleration, road grade climbing, braking deceleration, and braking force distribution on the front and rear wheels.

► [Vehicle Energy Storage: Batteries](#) gives an overview of batteries for vehicle applications and discusses the research and development roadmap of next-generation batteries for vehicle applications. It describes how the battery is used in EV and HEV and what the vehicle requirements on it are. It also presents the basic state of the art on the available batteries today.

► [Battery Technologies](#) describes the internal working of the electrochemical battery, the relationship between the internal resistance of the battery cell, maximum power output, energy efficiency in charging and discharging, and the issue of cell heating. A review of the characteristics of lead acid, nickel metal hydride, lithium-ion batteries, and super capacitors are then presented.

► [Energy storage: Ultracapacitor](#) describes the new technology of ultracapacitors for high-power energy storage and delivery. This is another approach to providing energy storage for hybrid vehicles. Ultracapacitors can be used either alone or in combination with batteries. In this approach, the batteries can be optimized for energy density and cycle life, and ultracapacitors can provide the power both for acceleration and regenerative braking.

► [Hybrid Energy Storage Systems for Vehicle Applications](#) describes the possibility of composing a high-energy, high-power hybrid energy storage system for vehicle applications, based on the analysis of inherent characteristics of different energy storage methods. The basic components in this system are chemical batteries, ultracapacitors, and flywheels (mechanical batteries). The objective is to fully utilize the advantages and suppress the disadvantages of each component to compose a compact energy storage system with high energy capacity and high power capability.

► [Batteries, Battery Management, and Battery Charging Technology](#) describes issues related battery charging and battery management. Battery management is one of the important engineering challenges of using batteries in vehicles. It starts with an overview of different types of rechargeable battery chemistries. Then, it discusses circuits for charging and managing vehicle batteries.

► [Vehicle Traction Motors](#) describes the role of the electric motor in EV and HEV drivetrains. The requirements that the vehicle put on the electric motor are explained. This is used as the basis to present the various traction motor drive technologies.

► [AC Machines: Permanent Magnet Synchronous and Induction Machines](#) describes the most common electric machines, motors, and generators that are used in EV and HEV drivetrains. It presents AC machines including induction machine (IM) and permanent magnet synchronous machine (PMSM). It is shown that both the IM and PMSM take advantage of the rotating magnetic field that is produced by their three-phase stator windings.

► [Switched Reluctance Motor Drives for Propulsion and Regenerative Braking in EV and HEV](#) describes a new basis for the selection of vehicle traction motor technology. The concept of extended speed

range at constant power and its benefit in vehicle traction application is introduced. On this basis, the modern switched reluctance motor drive technology is shown to be very attractive for future EV and HEV applications. In addition to having a wide speed range, these motors have a rugged and low-cost structure, reliable drive power converter topology, and high efficiency over a wide range of speed.

► **Regenerative Braking** describes the important technology of regenerative braking in EV and HEV drivetrains. Regenerative braking is responsible for most of the battery energy and fuel savings in these vehicles. This entry presents the principles of regenerative braking operation, design, and its hardware.

► **Hybrid Electric and Hydraulic Technology Applications in Off-Road Vehicles** describes a special class of HEV that are more suitable for off-road and heavy-duty applications. These include military vehicles, construction and mining vehicles, and similar industrial and special-purpose heavy uses. The special use of hydraulics in the drive train as part of the hybridization of the vehicle is of particular interest in this entry.

► **Plug-in Hybrid Electric Vehicles** describes a special class HEV that is called plug-in HEV (PHEV). It is shown that the PHEV has certain advantages over pure EV and HEV, such as (1) displacement of significant amount petroleum fuel by electric energy that comes from utility grid, (2) longer range than a pure EV, (3) no fuel consumption and emission during pure EV mode, and (4) lower overall fuel consumption and emission.

► **PHEVs and BEVs in Coupled Power and Transportation Systems** describes the role of utility electric power systems in the EV and HEV transportation system. The impact of plug-in HEV and battery-operated EV (PHEV/BEV) transportation systems on the utility power systems is discussed. The mutual benefits of charging the EV/HEV from the utility grid and some of their drawbacks are addressed.

► **Internal Combustion Engines, Alternative Fuels** for describes the technical implications of alternative fuels on vehicle engines, their performance and emissions characteristics. The characteristics of the most dominant alternative fuels are defined and compared to conventional gasoline and diesel fuel. Of particular interests are the renewable alternative fuels.

► **Vehicle Biofuels** describes the nature and origins of liquid fuels for vehicle engines that are derived from biomass. Liquid biofuels are the most valuable because they are easily stored, have a high energy density, and are readily metered into engines. It is shown that biofuels are infinitely renewable provided appropriate agricultural practices are followed.

► **Sustainable Vehicle Fuels, Well-to-Wheel Analysis** describes the important stages, results, and key issues associated with the well-to-wheel stages of biofuels, hydrogen, and electricity as alternative transportation fuels. The importance of comparing alternative vehicle engine fuels with the conventional fossil fuels is presented from the fundamental point of view.

► **Fuel Cell Powered HEV Design and Control** describes different types of fuel cells and their electrolytes, sources of fuels for fuel cells, those applicable to the transportation sector, and the hybrid vehicle architecture. The ICE hybrid is also presented as a base model, since its operation is well understood by the general public.

Electrical Transmission Systems and Smart Grids, Introduction

MIROSLAV M. BEGOVIC

School of Electrical and Computer Engineering,
Georgia Institute of Technology, Atlanta, GA, USA

Transmission systems represent the backbone of the electric energy. They support transport of electric energy from large producers (power plants) to the load centers (residential areas, manufacturing facilities, business centers or a combination thereof). Those networks are probably among the largest human-made engineering systems – the transmission network in the United States covers over 300,000 km of lines and is served by 500 companies (electric utilities).

In contemporary power systems, the notion of net energy producers and net users is increasingly blurred as most economic generator capacities are of smaller size and can be (and often are) installed near users' locations – example of small photovoltaic generators or wind farms which can be installed on the roofs of

residential homes or in their vicinity illustrates that alternative, popular for many reasons (little or no maintenance needed, decreasing cost of electricity generated by such small generators, many of which are based on various renewable energy sources, reduction of congestion which is the result of carrying large amounts of power across vast distances, reduction of transmission losses, regulatory and policy-driven economic incentives for small owners of generators, reduction of carbon footprint, and enhancement of sustainability of such solutions, etc.) The need for electric energy systems is not only to run the equipment in manufacturing facilities or appliances in residential homes, but also to interact in various ways with other supporting infrastructures (water, gas, transportation, information, etc.) Those infrastructures are interdependent on one another and their efficient and reliable operation requires a thorough understanding of those interdependencies and adequate planning and operational support (see ► [Energy and Water Interdependence, and Their Implications for Urban Areas](#)).

As electric energy travels across waveguides (conductors), it incurs losses due to dissipation of current across the resistances of imperfect conductors. High-voltage overhead conductors do not need insulation. The conductor material is nearly always made of an aluminum alloy, which typically forms several strands and often is reinforced with steel strands for mechanical strength. Copper was more popular conductor choice in the past, but substantially lower weight and cost of aluminum and its only marginally inferior electrical performance have been the reasons for the current dominance of aluminum-based conductors in the transmission networks. As large amounts of power being transferred across the lines may incur considerable losses in transmission, typically such bulk transfers are performed at higher voltages, which require smaller currents (losses in the conductors are proportional to the square of the current, which means that operating the transmission line at double the voltage incurs only about 25% of the losses produced while operating at lower voltage). It is the need for changing the operating voltage levels as a function of the energy throughput that has forced design transition from the original DC transmission (introduced by Thomas Edison) into AC, originally

deployed by Nikola Tesla and George Westinghouse in the 1880s to transfer power from the power plant at Niagara Falls, and nowadays used almost everywhere. Today, the highest operating AC transmission voltages can be up to 500 kV and even 800 kV. Ironically, when need for high power transfers calls for operation at voltages higher than 800 kV, it is done via DC transmission and with use of the large HVDC converter stations. The reason is inductances of the large overhead transmission lines, which at the highest voltages ultimately choke the efficient transmission of electric energy.

Transmission voltages are usually considered to be 110 kV and above. Lower voltages such as 66 kV and 33 kV are commonly called sub-transmission voltages. Voltages less than 33 kV are mostly used for distribution. Design of distribution networks is driven by their traditional role as infrastructure for disseminating bulk electric energy to a large number of customers. Such fragmentation of delivery requires distribution networks to operate somewhat analogous to capillaries in a cardiovascular system, at smaller capacities (and lower voltages) and covering large areas of sparsely populated customers (in rural areas) or densely populated smaller areas (in modern urban settings where increasingly large part of the world population now resides). Traditional design is evolving of radial distribution network, consisting of feeders supplied from the substations which interconnect them to the bulk power transmission networks. Such simple configurations were enabled by unidirectional flows of energy and simple distribution hardware which was supporting it. Transformation of distribution system into the site of both consumption and (distributed) generation of electric energy, as well as increasing need for enhanced interconnectivity at the distribution level, is imposing need for different designs. Part of the contemporary distribution networks are likely to experience a transition to microgrids (more meshed and better controlled distribution networks which can be used flexibly as reconfigurable autonomous, or grid-connected infrastructure for distribution of electric energy).

The structure and function of electric substations is also changing rapidly. Substations represent a vital part of the power grid infrastructure with many important functions (ability to reconfigure the system topology,

isolate equipment for maintenance and repair, monitor and communicate various system parameters and electrical variables to the control center or elsewhere, actuate control actions or protective relaying decisions, etc.) Substation automation represents one of the fastest evolving parts of the modern (smart) grids and will continue its evolution to keep up with the demands for more flexible and effective monitoring, control, and protection of power systems, both on the transmission and distribution side (see ► [Distribution Systems, Substations, and Integration of Distributed Generation](#)).

A growing part of the distribution networks, especially in developed countries, is being served by underground cables instead of overhead lines. There are many reasons in favor of such solutions (reliability, esthetics of the area where underground lines are installed, less vulnerability to the elements, etc.) and some against (considerably higher cost, faster pace of aging, especially due to moisture and impulse electrical stress, such as a consequence of lightning strikes in vicinity of the installations, and lack of effective diagnostic procedures to assess the operational status and effective remaining lifetime of the cables, more expensive repairs, etc.) Nevertheless, underground distribution (and in some places, underground transmission) represent a growing portion of the energy systems' assets, both in terms of importance and cost, and considerable care needs to be paid to their management and upkeep (see ► [Underground Cable Systems](#)).

Distributed generation (DG) can be defined as small-scale, dispersed, decentralized, and on-site electric energy systems. Currently, capacities of DGs vary typically in the range of several kW to hundreds of MW. As more DG penetrates the electric energy systems, more accurate and efficient system analysis algorithms are needed in order to analyze the impact of the DG system on various types of microgrids and distribution networks. Since DG can change the operation of the distribution system and interfere with its protection and control, electric power utilities are not motivated to interconnect customer-owned small generators to their distribution networks. Utilities tend to put nonutility generation under the extensive technical analysis. Conversely, the regulating authorities tend to act in favor of DG owners and support

that the interconnection be as easy and transparent as possible. Nevertheless, the favorable economic features of small-scale distributed generators, especially those using the renewable energy resources as input, will make them increasingly popular and much more widespread than they are at the moment of creation of this text. Even now, many countries (Ireland, Spain, Denmark, etc.) possess considerable renewable resources as part of their generation portfolios. At the time when wind generation of electricity is the fastest growing new generation technology (in terms of new installed capacity) and when the energy produced by renewable resources (including hydro plants) is already larger than energy obtained from the nuclear power plants (United States in 2011), engineering challenges of planning, operating, controlling, and protecting the new power grids are substantial and require major transformative changes (see ► [Renewable Generation, Integration of](#)). This may become even more important as some countries elect to gradually abandon conventional nuclear generation and transition to other, more sustainable, generation resources.

One of the fundamental constraints in the transmission and distribution of electricity is that, for the most part, electrical energy cannot be stored, and therefore must be generated whenever needed. Few exceptions to that limitation have been found and exploited. The biggest problem is that currently available storage options cannot effectively be used at utility-scale capacities (pumped hydro plants are the best known among them). In the interim, a large number of smaller capacity storage technologies have been developed and advanced (superconductive magnetic storage, flywheel, battery storage, etc.), mostly to find applications as uninterruptible power supplies for critically important (but relatively small) loads and rarely exceed the capacities needed to achieve similar effects in the bulk power networks. When a storage technology is developed and commercialized to operate at such large capacity levels economically, it will trigger a major revolution in power grid planning and operation.

A very advanced infrastructure of monitoring, control, and protection (see ► [Wide Area Monitoring, Protection and Control](#)) is required to enable real-time balancing between electric generation and energy demand due to the lack of effective storage options.

If generation and consumption of electric energy are not in balance or if the complex infrastructure of voltage control across the system is challenged by the heavy loading conditions and/or the aftermath of an unforeseen major disruption in system operation (such as the outage of a major piece of equipment), generation plants and transmission lines can shut down which may lead to a major blackout. Sometimes such blackouts develop spontaneously in a cascading chain of equipment outages caused by spreading of the overloads through the equipment system-wide as a consequence of an initial disturbance, which may be relatively minor in its initial effects. These series of events may occur in unanticipated sequence, and can be hard to foresee even with large computers. An unexpected contingency may cause an amplifying effect of larger sequential outages and progressively stress the system to the point when the disturbance can no longer be contained. In the domain of large power grids, “flapping of the butterfly wings” in certain places can literally produce a “storm” in others (see ► [Transmission Blackouts: Risk, Causes, and Mitigation](#)).

To reduce the risk of such failures, electric transmission networks are highly meshed and interconnected into regional, national, or continental wide networks, providing multiple redundant alternate paths for energy to flow when needed. Considerable effort is expended by electric utility companies to ensure that sufficient spare capacity and redundant pathways for energy transfer are always available to mitigate the consequences of even large multiple disruptions of the network. In order to maintain sufficient security margins under the threat of multiple unpredictable contingencies, well-coordinated plans for preemptive (slower, based on extensive optimization algorithms applied to coordinate generation and control system wide) and preventive (much faster, emergency control and protection) actions need to be developed so that the system does not descend into a blackout and necessitate lengthy and costly restoration procedures (see ► [Smart Grids, Distributed Control for](#)). Particular focus should be on determining the ability of system to survive extreme contingencies, triggered by very unlikely chains of events, but capable of propagating into costly widespread outages with long-term consequences for both consumers and the power companies.

All of the above functional characteristics and solutions describe what is commonly referred to as smart grid. Gradual application of emerging technologies for advanced power grid management and control/protection represents an effective transition to smart grid, which from the perspective of different authors and researchers may assume different characteristics, but in general shares the following functional properties:

- Ability to resiliently recover to the extent possible from the effects of damaging or disruptive disturbances (self-healing)
- Providing opportunities for consumer participation in energy management and demand response (often via advanced metering options which may provide additional support and information, both for the utility and the customer, during normal operation)
- Ability to respond to, cope with, and resiliently enhance itself against physical and cyber attacks
- Providing power quality for modern equipment anticipated to be needed in the future
- Accommodating all generation and storage options
- Enabling new products, services, and markets
- Optimizing assets and operating efficiently

The transition to smart grid may in some cases be spontaneously driven by obvious benefits and cost-effectiveness, in others may be supported by regulatory and policy actions (see ► [Sustainable Smart Grids, Emergence of a Policy Framework](#)).

Electricity Generation with Small Wind Turbines

G. J. W. VAN BUSSEL

Aerospace Engineering Section Wind Energy,
Delft University of Technology, Delft, The Netherlands

Article Outline

Glossary

Definition of the Subject

Introduction

The Feasibility of Small Wind Turbines

Small Wind Turbine Technology

Performance of Drag-Driven Wind Turbines
 Performance of Ducted Wind Turbines
 Performance of Lift-Driven VAWTs
 Control and Safety Systems
 Costs of Small Wind Turbines
 Bibliography

Glossary

COE Cost of Energy: A method to calculate the electricity costs taking into account investment and operation and maintenance costs.

Control and safety systems Systems to keep small wind turbines within their operational range and to bring them back into a normal operational mode.

Darrieus rotors A VAWT with curved rotor blades using a lift principle for its operation.

DAWT Diffuser-Augmented Wind Turbine or ducted wind turbine.

Drag Force in the direction of the incoming flow on a (streamlined) body.

HAWT Horizontal Axis Wind Turbine.

H-type VAWT A VAWT with rotor blades rotating along a cylindrical surface.

Lift Force perpendicular to the incoming flow on a (streamlined) body.

Savonius rotors “Crosswind” wind turbines with the main shaft perpendicular to the wind.

VAWT Vertical Axis Wind Turbine.

Definition of the Subject

Small wind systems can be seen in autonomous operation for pumping water, desalinization, and battery charging. A growing number of small wind turbines is connected to the grid, most of them behind the meter. This entry will focus on electricity-generating small wind turbines. Since the electricity production with wind turbines on buildings is a fairly new development with an interesting potential it will also be addressed. Safe and reliable operation of small wind turbines is technically challenging since they usually operate in more turbulent flows than large-scale wind turbines. A short description of the working principles of various types of small wind turbines is given; their safety and control systems are discussed as well as noise and vibrations. Furthermore, a treatment of the investment

costs and the cost of energy is provided. Finally the implications for locating wind turbines on top of buildings are treated.

Nontechnical barriers such as legislative and regulatory issues are not addressed, although they are very important for successful implementation of small wind turbines.

Introduction

There is a growing interest in the use of small wind turbines. The market is already significant and it is expected to grow at a similar rate as large-scale wind power. Small wind turbines are not only developed for remote applications in rural areas but also, to an increasing extent, for application in urban areas and on buildings. This entry will address small wind turbines for electricity generation. Other important applications, such as wind pumps for irrigation, drainage, and the provision for drinking water, are not addressed.

For further reading about wind pumping technology the reader is referred to Lysen [24] Kentfield [25] and de Jongh and Smulders [23].

In a strategy document issued by the American Wind Energy Association, a small wind turbine committee chaired by Bergey [2], a projected total of 75,000 MW wind power capacity by 2020 is provided by small- to medium-scale systems. With an average of 2 kW per turbine this would mean around 30–40 million units. Of this amount some 15,000 MW is foreseen at locations that are (also) suitable for rooftop mounting of wind turbines. A potential market of similar extent is present in the EU (Van Bussel 2005). Small wind turbines are also a very interesting option for small-scale energy supply at remote sites where no grid electricity is available and a grid extension would be too costly (e.g., for remote villages or homes). As such, they may have a great potential in developing countries. The most striking example is probably their dissemination among the nomads in Inner Mongolia (China), where over 100,000 small portable wind turbines (capacity about 100 W) are in use for “domestic” power supply. Such small off-the-grid wind turbines are often referred to as wind chargers [22].

Small wind turbines are not only developed for remote applications in rural areas but also, to an increasing extent, for application in urban areas and

on buildings. Yet the technology and the costs of small-scale wind deployment are not at the level of large-scale wind deployment. There is not a unique definition of what is regarded as a small wind turbine or a small wind system. The AWEA uses an upper limit of 100 kW and a maximum rotor diameter of 20 m [2] while the British Wind Energy Association [26] takes 15 kW and an annual production of 50,000 kWh as the upper limit without referring to any rotor size. Others [28] take a maximum swept area of 50 m² as the criterion. In this entry, the viewpoint is taken that a small wind turbine system is a system that is able to provide (part of the) electricity consumption for a local (private) customer.

Wind turbine deployment in the built environment is inherently more complicated than in rural, open spaces. In successfully exploiting wind energy in the built environment, three major issues can be identified: wind resource assessment and wind characterization around buildings, structural integration of wind turbines with buildings, and special urban wind turbine design requirements.

Safe and reliable deployment of wind turbines in the highly turbulent built environment is specialized and

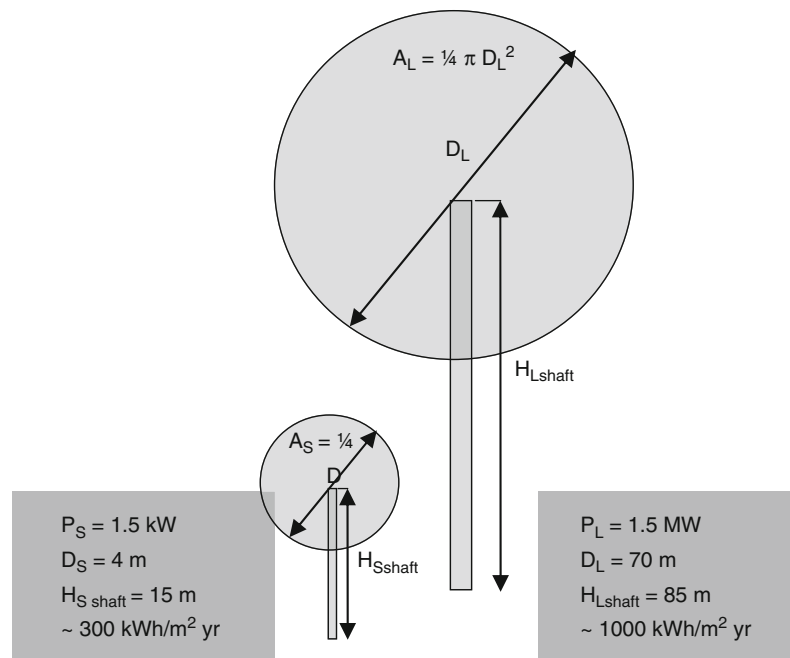
technically challenging. In almost all situations, the use of existing (small) wind turbines will be problematic due to the fact that they are not adapted to the complex wind environment and to additional building-related requirements. Some of these requirements are the severe noise restrictions and the ability to match the structural and esthetical integrity of buildings. This means that new ways need to be found in the design of new wind turbines that do fit the requirements for the built environment.

Nontechnical barriers such as economics, bureaucratic, and regulatory issues will not be addressed, although they are very important for successful implementation of wind turbines in the built environment.

The Feasibility of Small Wind Turbines

This section shows a comparison between large and small wind turbines. The comparison is performed in order to show if there is space for small wind turbines in the built environment from an economical point of view.

Consider a large wind turbine with a *rotor diameter* D of 70 m and a *swept rotor area* A_S of 3,848 m² (Fig. 1).



Electricity Generation with Small Wind Turbines. Figure 1
Comparison of large and small wind turbines

The height from the ground to the nacelle and main shaft H_{shaft} is 85 m. The wind turbine yield is assumed to be 800–1,100 kWh/m²/year, evidently depended on its location. This means that around 3,000–4,500 MWh/year can be produced with a wind turbine of this size [8].

In the West European situation, this means that a few (3–10) m² of the swept surface suffices to meet the average electricity demand of one family.

Now take a small wind turbine with a diameter of 4 m at a 15-m tall mast. The swept area is then 12.57 m². With an average yield of 300 kWh/m²/year (feasible in a very good wind climate: an average of at least 6 m/s at hub height) this is sufficient for one family (3,770 kWh/year). It thus takes 1,000 of these relatively small wind turbines to provide the same amount of electricity.

In normal utility grid operation it will be very difficult to be competitive in terms of the costs per kilowatt hour. Either at remote locations where conventional electricity is very expensive and building large wind turbines is not possible or in situations where it is allowed to generate electricity behind the meter, since consumer price of electricity may easily be three times the production costs, small wind turbine may become a feasible solution.

The consumer's price of electricity was 0.22€/kWh in 2008 in the Netherlands. The Dutch consumer's price is roughly built up out of generation of electricity (30%), taxes (40% incl. carbon tax), and the distribution and administration costs (30%). This consumer's price may give an indication for the level of the cost of electricity which might make small wind turbines economically viable.

Small Wind Turbine Technology

Most of the modern small wind turbines typically have an upwind propeller type rotor, often directly coupled (without a gearbox) to an electrical generator (Fig. 2a). Such wind turbines are usually referred to as horizontal axis wind turbines (HAWTs). An increasing amount of small wind turbines pop up on the market with unconventional rotor topologies, such as Savonius rotors (Fig. 2b), Darrieus, and H-type vertical axis rotors (VAWTs) (Fig. 2c and d), as well as horizontal axis wind turbine rotors equipped with a duct (DAWTs) (Fig. 4a).

Vertical axis wind turbine can be further distinguished through their principle of operation. Some are driven by drag differences between the forward moving and the retracting blade; others use the lift principle, similar to that of “propeller”-type HAWTs.

Small wind turbines have mechanisms to limit the rotor speed and the forces on the rotor at high wind speeds. The rotor speed is usually allowed to vary with the wind speed. In the past, small wind turbines were often equipped with a direct current (DC) generator. Nowadays almost all machines are equipped with alternating current (AC) generators, most of them using Permanent Magnets (PM generators) which supply an electric output with variable voltage and variable frequency. An inverter is used to convert this variable speed AC into either fixed frequency AC to comply with the local grid convention, or to DC when the wind turbine operates in an off-the-grid mode using batteries for storage.

Since the principle of operation is an important issue with respect to the aerodynamic efficiency of the machine it is important to address the aerodynamic theory of these machines shortly.

Performance of Drag-Driven Wind Turbines

The conversion of wind energy into rotational energy of drag-driven wind turbines is based on the drag difference of rotating bluff bodies (as in Fig. 3).

An anemometer like drag-driven wind turbine is depicted. It consists of bluff bodies with different pressure drag induced by their shape. This difference in drag causes the turning moment.

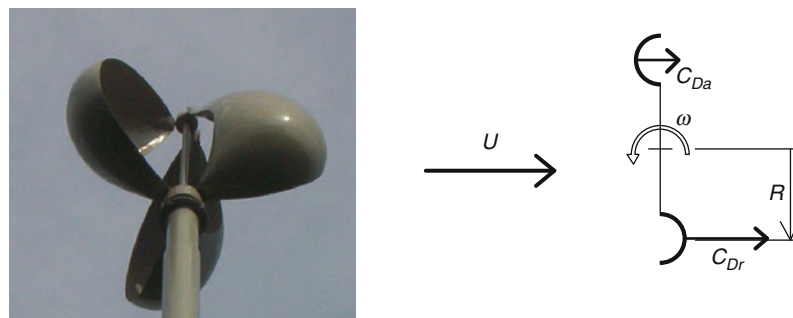
The “tip speed ratio” λ is defined as $\lambda = \frac{\omega R}{U}$ where ω is the rotational speed, R the radius of the rotor, and U the incoming wind speed.

The aerodynamic efficiency defined as $C_P = \frac{P}{\frac{1}{2}\rho U^3 A}$ where A is the total area swept by the wind turbine rotor projected on a plane perpendicular to the wind direction and P is the (aerodynamic) power extracted from the wind. The nondimensional coefficient C_P is called the power coefficient and is a measure of the (aerodynamic) efficiency of the machine. Where the theoretical maximum power coefficient of lift-driven wind turbines is $C_P = \frac{16}{27}$, the maximum attainable C_P of a drag-driven wind turbine can be approximated by $C_P \approx \frac{4}{27} C_{Dr}$, where C_{Dr} is the drag



Electricity Generation with Small Wind Turbines. Figure 2

(a) Small HAWT (b) drag-driven VAWT (c) small lift-driven VAWT (d) H-type lift-driven VAWT



Electricity Generation with Small Wind Turbines. Figure 3

Left: AES drag-driven VAWT, right: aerodynamic working principle

coefficient of the retreating (moving with the wind) rotor blade. Here it is assumed that the drag coefficient of the advancing blade is zero: $C_{Da} = 0$, where this is only the case when the advancing blade is completely sheltered from the wind.

Drag coefficients of bluff bodies have maximum values up to 2; therefore the optimal aerodynamic efficiency of a drag-driven wind turbine will be up to $\frac{8}{27} \approx 0.3$. In practice, drag-driven wind turbines usually have power coefficients below 0.1. This is in contrast to lift-driven machines which are able to achieve significant larger power conversion efficiencies. Large-scale HAWTs do have C_p values ranging as high as 0.5 where small-scale wind turbines usually level off at values around $C_p = 0.35$.

Performance of Ducted Wind Turbines

The performance of wind turbines equipped with a duct may be larger than “bare” wind turbines. Figure 4 shows an example of a small DAWT.

As can be seen from the schematic on the right-hand side of Fig. 4, the duct is able to induce a larger mass flow through the rotor: the yellow and red colored areas indicate areas where increased velocities are found. This increased mass flow gives rise to larger performances of a propeller-type rotor located at that position. The performance of well-designed ducted wind turbines compares well with the performance of a “bare” HAWT when referencing to the exit diameter of the DAWT. In other words, if the exit diameter of a DAWT is the same as the rotor diameter of a HAWT, they are able to produce similar power at the same wind

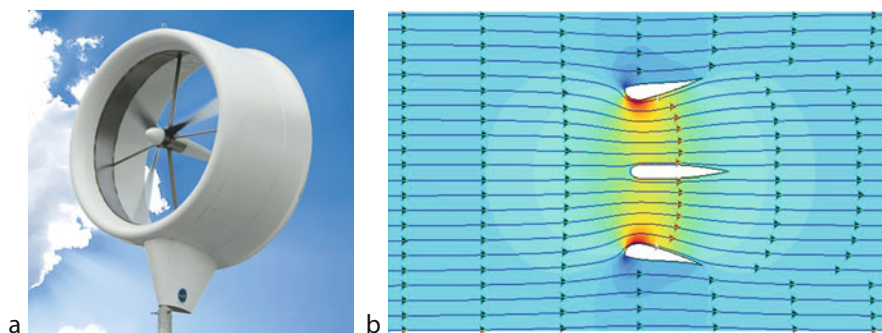
speed [6]. The duct of a DAWT acts as a tail vane and keeps the rotor oriented toward the wind. The possible largest advantages of a DAWT are the possibility to incorporate noise damping material into the duct and the fact that the outer area of the duct can be used for commercial and/or styling purposes. This, especially, holds for DAWTs operating on top of company buildings.

Performance of Lift-Driven VAWTs

The aerodynamic efficiency of a wind turbine depends upon other things on the way the wind energy is converted into mechanical power.

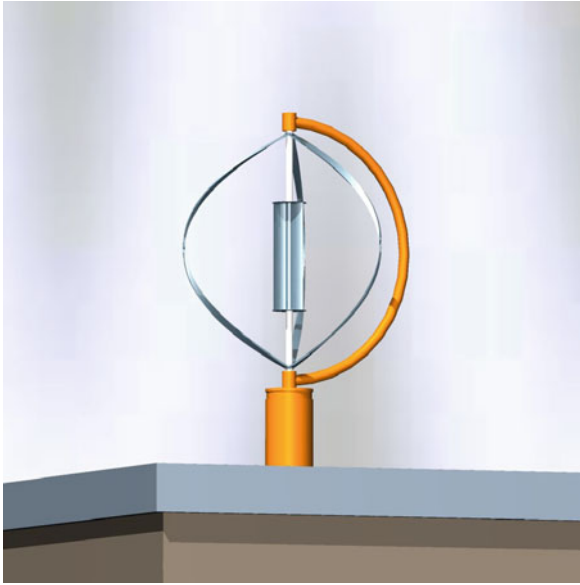
A high aerodynamic efficiency can be achieved when the working principle is based upon lift forces. Such wind turbines make use of airfoil-shaped blades in order to generate their driving force.

Lift-based VAWTs usually run fast and have slender blades (see Fig. 5), just like modern large-scale HAWTs. This is in contrast to drag-driven VAWTs, which achieve a significantly lower aerodynamic efficiency and use a large amount of rotor material. The Darrieus wind turbine (DWT) is the most widely known example of a lift-driven VAWT. Its working principle is depicted in Fig. 6. The energy extraction takes place during the cross-flow phases of the blades, that is, when the blades move perpendicular to the oncoming wind. With respect to a modern HAWT the Darrieus wind turbine has a somewhat lower efficiency (about 10%) in an ideal undisturbed low-turbulence wind environment. The main reason for this is that the rotor blades do not contribute to generating torque



Electricity Generation with Small Wind Turbines. Figure 4

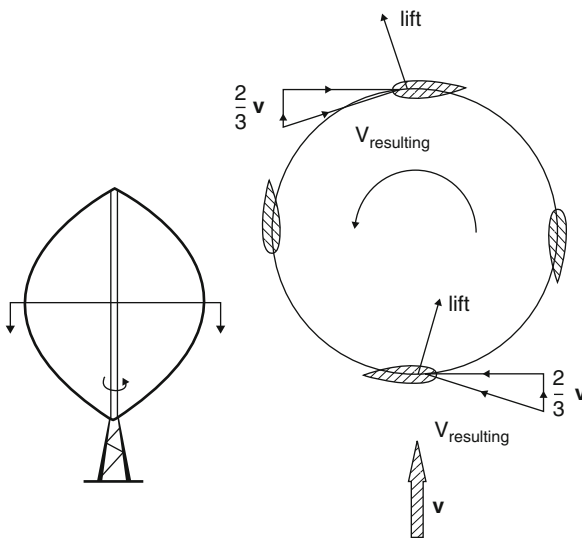
(a) small Donqi[®] DAWT wind turbine (b) aerodynamic operation principle of a duct around a DAWT



Electricity Generation with Small Wind Turbines.

Figure 5

Globuan, a modern Darrieus wind turbine located on a flat roof



Electricity Generation with Small Wind Turbines.

Figure 6

Aerodynamic principle of operation of the Darrieus rotor

when they are at the retreating and at the advancing phases of the revolution. On the contrary, then they generate drag only which has a detrimental effect on the performance.

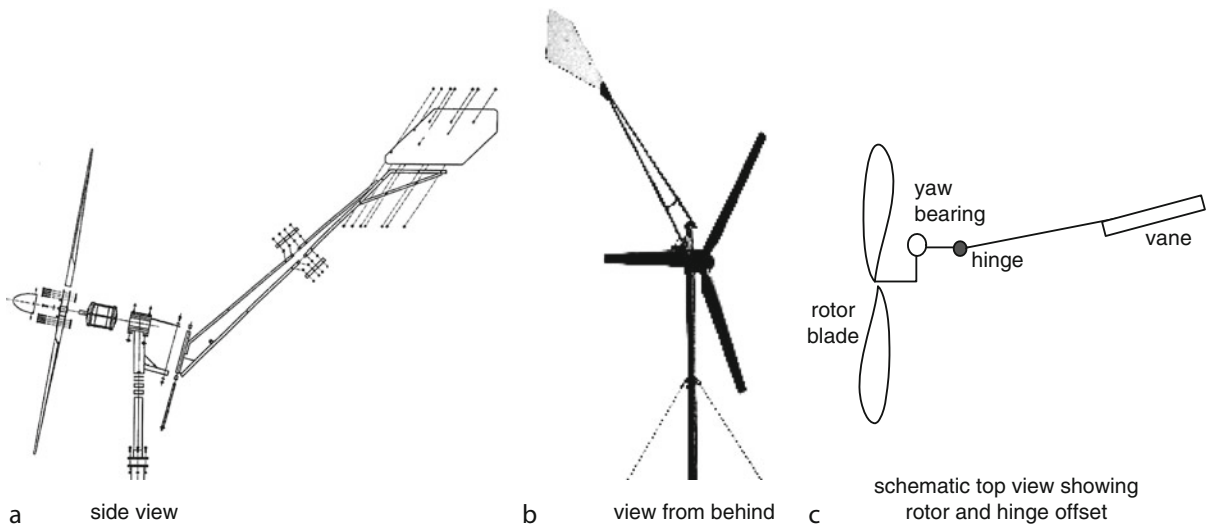
This also gives rise to a fluctuating torque. Novel small VAWTs have mitigated these torque fluctuations by introducing a blade sweep. In Fig. 17, examples are shown of VAWTs with blade sweep. A second, large advantage of blade sweep is the significant reduction of noise. Normally high tip speed ratio (slender blade) Darrieus-type VAWT machines are very noisy, where the noise generated resembles the noise of an ascending helicopter. Another disadvantage of a lift-driven VAWT is its incapability to deliver a starting torque. This means that auxiliary systems are needed to get the machine running. Sometimes a small Savonius rotor is used, as in Fig. 5, but that has the disadvantage that the aerodynamic performance of the total system is influenced negatively in the normal operation mode. For this reason a start-up procedure is often implemented in which the generator of the wind turbine is temporarily used as a motor, to acquire a sufficient rotational speed.

Control and Safety Systems

Each wind turbine needs a safety system to take action when the machine is getting out of its normal mode of operation. Also control systems are usually incorporated in the design. Control systems are used for maximizing energy production while keeping the forces on the system within limits. Control and safety systems can be implemented using the same wind turbine components, but in terms of hierarchy the safety system should always be able to overrule the control system.

Figure 7 shows a commonly used control and safety set up for a small HAWT. The machine is equipped with a tail vane. The first function of this tail vane is to keep the rotor oriented toward the wind. The tail vane of this system, however, is connected with a hinge to the body of the wind turbine. The tail vane hinge is tilted with respect to the direction of gravity (Fig. 7a). This causes the tail vane to prefer a “lowest” position due to the weight of the tail vane. For zero and low wind speeds the tail vane will be in this position keeping the rotor perpendicular to the wind and as such maximize the energy production for low wind speeds.

The rotor shaft, connected to the body of the turbine, has an offset with respect to the yaw bearing on top of the tower. Thrust on the rotor will initiate a yaw moment turning the rotor out of the wind. The



Electricity Generation with Small Wind Turbines. Figure 7
Fortis Montana® control and safety system

tail vane however prevents yawing. This is possible because the hinge of the tail vane has an offset and hence it can develop a yawing moment counteracting the yaw moment due to thrust on the rotor. But when the thrust on the rotor is getting too large, the aerodynamic force on the tail vane will cause the vane to rotate around its tilted hinge against gravity forces. This reduces the aerodynamic force on the tail vane and as a consequence, the thrust on the rotor will initiate a yawing movement leveling off the rotor thrust.

So in this hinged tail vane set up, the tilted hinged vane both has a control function as well as a safety function, since very large thrust forces are prevented by yawing the rotor out of the wind.

Apart from the tilted hinged vane other control and safety systems are found at small wind turbines.

Figure 8 shows two examples of safety systems activated by over speeding of the rotor. The ESI® wind turbine is equipped with tip brakes which are kept in place by springs in normal operation. Another commonly applied system uses centrifugal governors causing the blades to pitch toward feather when the rotational speed becomes too high.

Small wind turbines equipped with a permanent magnet direct-driven generator also may have a safety system through applying short circuit in the electric windings, causing the machine to stop.

This is a safety system that is also often applied in VAWTs since yawing the machine out of the wind and/or pitching the rotor blades are usually not an option in vertical axis designs.

Noise and Vibrations

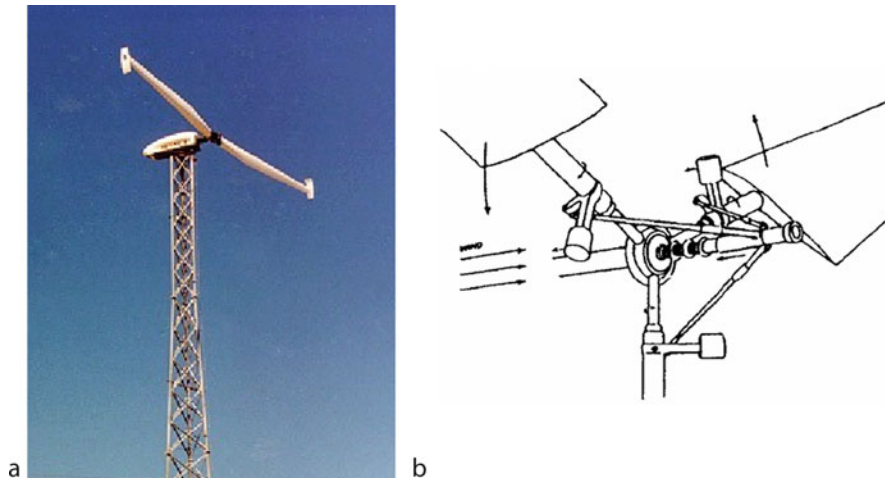
For small wind turbines noise and vibration can be critical issues. Usually small wind turbines operate in the vicinity of people, which imply that noise levels should be within acceptable (legislation) levels.

Vibrations tend to reduce lifetime, but are a vital issue for roof top locations since they might give yield to induced noise inside the building.

Noise Background noise level in rural areas starts at 25 dB(A) [9] at a 10 min averaged wind speed of 4 m/s and increases with wind velocity (see Fig. 9).

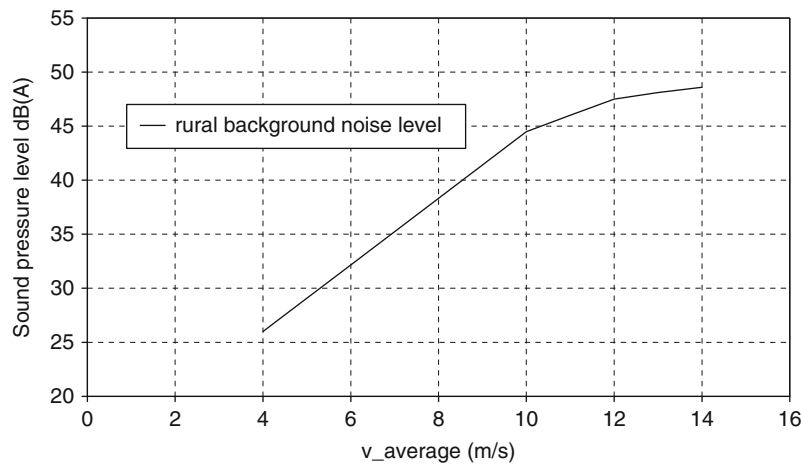
The noise level in a built environment depends on local noise production, for instance, traffic. Due to an increased human activity noise in the built environment will be higher compared to rural areas.

When the wind turbine is modeled as a point source with hemispherical spreading, the noise shows a reduction of 6 dB(A) per doubling of distance to the source (Fig. 10). The distance from the source to the location of noise perception is important, and the sound produced by the wind turbine should ideally be at a level beneath or at least equal to the background level.



Electricity Generation with Small Wind Turbines. Figure 8

(a) ESI® wind turbine with tip brakes (b) passive pitch system using a centrifugal governor, reducing rotor thrust at high rotational speeds (e.g., [22])



Electricity Generation with Small Wind Turbines. Figure 9

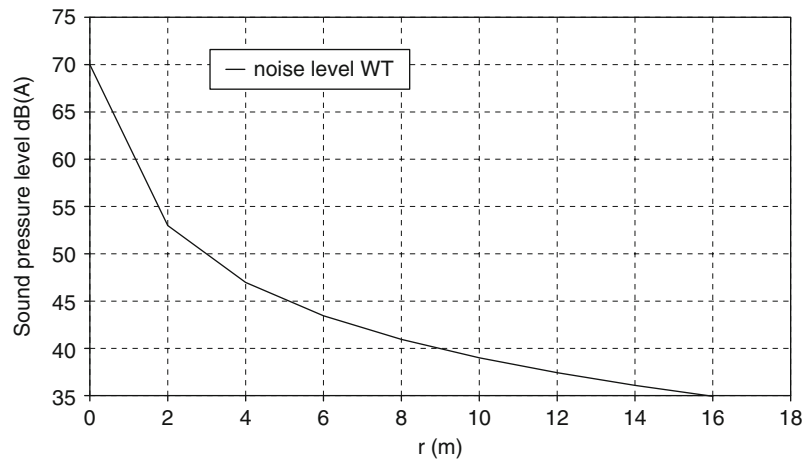
Prevailing background noise level in rural areas at an average 10 min [14]

An obvious approach to reduce aerodynamic noise emission is to lower the rotational speed of the wind turbine. Here a big advantage of drag-driven VAWTs comes in because of their inherently low rotational speed and hence noise production.

Lift-driven VAWTs have a big disadvantage with respect to noise production. Early (large-scale) Darrieus machines were notorious noise producers, where the cause of the noise production was the unsteady aerodynamic loading of the blades during a

revolution. Mitigation of that noise source has taken place in modern small-scale VAWTs either by reducing the tip speed significantly (e.g., Fig. 2d) or by introducing a blade sweep (see Fig. 17c).

Vibrations Vibrations occurring in structures should be minimized since they cause fatigue damage. Proper design of the components of a small wind turbine and proper choice of tower and foundation should result into low vibrations. Crucial in this respect is the tuning



Electricity Generation with Small Wind Turbines. Figure 10

Wind turbine noise level at distance r for a point source noise emission of 70 dB(A) [14]

of the natural frequencies of the tower. They should be outside the rotational frequency/frequencies (and frequencies equal to n times the rotational frequency where n is the number of blades).

Compared to large machines, small wind turbines operate at a much larger rotational frequency ω . The reason is that the wind U does not scale with the size and since a rotor is designed to operate at a given tip speed ratio λ where $\lambda = \frac{\omega R}{U}$, the ω must be larger when the radius R goes down. This is a potential noise hazard since this might mean that vibration frequencies are getting in the audible range (>20 Hz) and as such may contribute to noise production.

Wind turbines may generate vibrations at frequencies starting at the rotational speed and the higher frequencies $i\omega$ where i is an integer, n is the number of blades, and ω is the rotational speed of the wind turbine. Vibration of the nacelle is often a consequence of mass and/or aerodynamic unbalance of the rotor and can thus be avoided by proper design and maintenance. But even well-balanced blades moving through turbulent structures in the incoming wind may cause vibration frequencies as well. Compared to large machines, small wind turbines operate at a much larger rotational frequency ω .

Lift-driven VAWTs will induce frequencies in the building at ω (mass unbalance) and $2i\omega$ (aerodynamic vibrations) since during one revolution a blade passes turbulent structures twice: first at the windward side of the VAWT and again at the leeward side of the VAWT.

Consequences for the operating wind turbines on buildings are evidently that operating at the natural frequencies of parts of the building, such as floors and walls should be avoided. The fundamental frequency of the building itself, which can be approximated by $f_e = \frac{46}{h_{geb}}$ according to Mertens [14], is usually well below the operating frequencies of the wind turbine.

Comparison of Different Types of Wind Turbines

The comparison of the wind turbines based upon their rotor topologies is provided in Table 1 below. The different systems are ranked according to their aerodynamic efficiencies. But this is evidently only one of the aspects that have to be taken into account. Auxiliary systems such as systems to yaw the machine in and out of the wind as well as control and safety systems need to be taken into account as well. Furthermore also the costs of the total system will play a role in considering what may be the best option.

Costs of Small Wind Turbines

The costs of small wind turbines are of evident importance. At first, the production of a given machine should always be considered in regard to the investment costs and the operational costs, and second, an assessment of the real total costs is important for making an economic trade-off between the choice of different (renewable) energy sources.

Electricity Generation with Small Wind Turbines. Table 1 Comparison of different small wind turbine systems based upon their rotor topology

	DAWTs (Diffuser-Augmented Wind Turbines)	HAWTs (Horizontal Axis Wind Turbines)	Lift-driven VAWTs (Vertical Axis Wind Turbines)	Drag-driven VAWTs (Vertical Axis Wind Turbines)
Advantages	Aerodynamically most efficient	Aerodynamically most efficient	Aerodynamically efficient	
			Wind direction insensitive	Wind direction insensitive
				Turbulence insensitive
				No noise
		Well proven	Easy access to generator	Easy access to generator
Disadvantages	Large and expensive duct needed	Most widely used		Well proven
				Robust
		Wind direction change sensitive	Needs start-up aid	Aerodynamically inefficient
		Tail vane/yaw system needed		Large amount of rotor material
		Does not perform well in highly fluctuating winds		

Cost of Energy

The *cost of energy* (COE) is calculated with a recommended procedure of the International Energy Agency (IAE) (e.g., [27]).

$$\text{COE} = \frac{\text{CRF} \cdot I}{\text{APE}} + \frac{\text{TOM}}{\text{APE}}$$

where:

COE: *Cost of energy* [€/kWh]

I *Investment*: wind turbine costs [€]

CRF *Capital Recovery Factor*: yearly interest [%/year], depending on *interest rate* *i* and *economical lifetime* *n*:

$$\text{CRF} = \frac{i \cdot (1 + i)^n}{(1 + i)^n - 1}$$

APE: *annually produced energy* [kWh]

TOM: *total yearly operation & maintenance costs* [€]

When an interest rate of 6% per year is assumed as well as an economical lifetime of 15 years [28], this

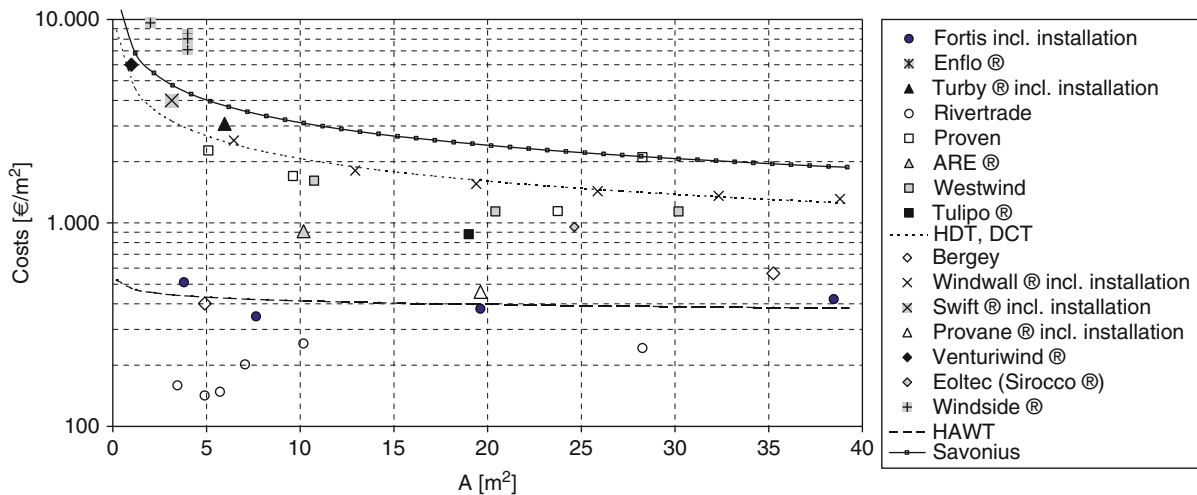
results in a CRF of 0.103. The investment costs of main components of a small wind turbine will be discussed below.

Wind Turbine Costs

Information of wind turbine component costs has been gathered extensively by Van Beveren [4]. Wind turbine costs and component costs are determined based upon the values in the year 2006 and are valid for wind turbines with a nominal power up to 10 kW. Costs are given as costs per swept rotor area, since this is the best ratio for comparison. Wind turbine costs could also be divided by its nominal power *P*, but since this value is not unambiguous, often determined by wind turbine manufacturers themselves and seldom overestimated, it is not an objective measure. Note the vertical logarithmic scale. If wind turbine costs include costs for installation, this is indicated by the label.

The trend lines are shown in the Fig. 11.

Costs of HAWTs, the most common small wind turbine design as function of the swept rotor area,



Electricity Generation with Small Wind Turbines. Figure 11

Year 2006 wind turbine costs and swept rotor area as specified by various manufacturers

comply roughly with $C_{\text{HAWT}} = 500 \cdot A^{0.9}$, where the costs of DAWT and lift-driven DAWT can be described by $C_{\text{DAWT}} = 4000 \cdot A^{0.6}$. The cost data of Windside® turbines, a typical drag-driven VAWT, show even higher costs and are described by $C_{\text{Savonius}} = 6000 \cdot A^{0.6}$. These trend lines are also shown in Fig. 11.

As can be seen from this figure, the standard HAWT has by far the lowest cost to area ratio, though there is also a large spread in the values. The Proven machines turn out to be almost twice as expensive as the Bergey machines, while both are respected manufacturers with a relatively large production. What can be seen is that the small HAWT costs curve is only slightly dependent upon swept area (diameter in this case). The other types of wind turbines show a different trend. It should, however, be realized that the amount of data available for these machines is limited, as well as the number of produced machines.

Tower Costs

A distinction is made between lattice towers, steel guyed masts, supported tube towers, and free standing tube towers (see Fig. 12 taken from [4]). The costs of these towers as shown in the figure are for tower heights of 5–37 m.

The costs of the tubular towers of the lift-driven VAWT Turby® deviate substantially from the other designs. The most probable reason is that these are

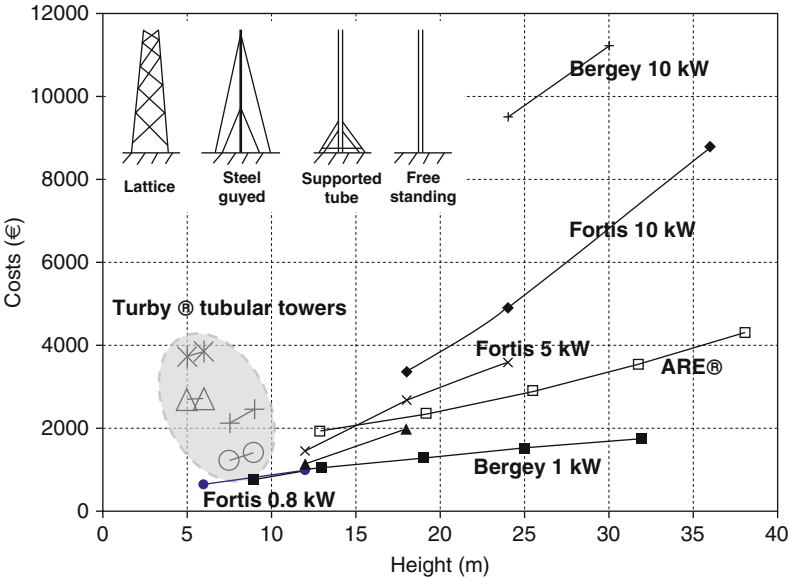
free-standing towers, where guyed towers are mostly used for Darrieus-type machines. Turby® though was designed for roof top application where guy wire connection may cause substantial additional installation costs, which are avoided in the current design.

Costs of Foundation and Installation

Costs of foundation and installation are based upon the WINEUR report [28] and further data of the manufacturers, a.o Fortis and Proven. Some towers need a foundation of concrete, others use a base plate with guyed wires attached to separate anchors. Evidently the installation and foundation costs depend very much on location, type of tower, and ground surface. Turby® has designed a special support structure for the installation upon a building. Of all cost elements of a small wind turbine these are costs which are most probably best suited for DIY activities, in order to bring the total investment down. A typical estimate for the professional costs of foundation and installation of small wind turbines is between 10% and 20% of the total investment.

Operation and Maintenance Costs

On top of the investment cost the Operation and maintenance (O&M) costs contribute to the total cost of energy. In the COE calculation, these are represented



Electricity Generation with Small Wind Turbines. Figure 12
Tower costs as function of tower height

in the TOM component. Total O&M (TOM) costs are an addition of *inspection and maintenance costs* C_{IM} and the *cost of replaced parts* C_{RE} and thus: $TOM = C_{IM} + C_{RE}$.

Inspection and Maintenance Visit Costs Most wind turbine manufacturers advise to have the wind turbine checked at least twice a year. The costs associated are low if the wind turbine is inspected by the user. Normally inspection of the machine can be done within 1.5 h for an average-sized wind turbine with a swept rotor area of 10 m². Larger machines may take more time, mainly caused by the extra time and effort to access the nacelle at larger heights.

Replacement Costs The replacement costs are estimated with the *failure rate* of the wind turbine parts. These failure rates are based on an assessment of Van Beveren [4] covering a period of 10 operational years of small wind turbines and values for large-scale wind turbines of Van Bussel [3] (Table 2).

These values indicate that once every 1–3 years a part of a wind turbine fails and needs replacement. This is significantly better than for large-scale wind

Electricity Generation with Small Wind Turbines.

Table 2 Estimate of failure rates of small wind turbine parts

No.	Wind turbine parts	Failures/year
1	Rotor blade	0.02–0.03
1S	Rotor blade Savonius	0.001–0.01
2	Shaft	0.001
3	Hub	0.001–0.05
4	Bearings	0.1–0.2
5	Generator	0.02–0.04
6	Power regulator	0.03–0.06
7	Inverter	0.1–0.2
8	Frame	0.002–0.004
9	Tail	0.03–0.06
10	Tower	0.01–0.02
11	Foundation	0.001
12	Cables	0.001
13	Anemometer and data logger	0.030
14	Lightning protector	0.001

turbines, maybe because smaller wind turbines allow for a larger safety margin in the design.

Impact of O&M Costs on Energy Production

Downtime due to inspection, maintenance, and repair is not included in the TOM costs, but is found in the APE part of the COE calculation. Failure rates and inspection times, however, indicate that the availability (defined by the ability to produce if wind conditions are within the operational window) of small systems is larger than for large-scale systems. Since modern large-scale wind turbines show values beyond 98% downtime, production loss for small systems should not be an issue. But this is of course subject to the logistics of spare part handling.

Indications of APE cannot be given, since they strongly depend upon the local wind climate. In general terms, a kWh production of more than 1,000 kWh/m² per year swept rotor surface is feasible for large-scale wind turbines in good wind climates. Small-scale wind turbines are usually located at less favorable sites and operate at lower heights. Then a production level of 300 kWh/m² per year is considered to be feasible on a good location.

Application of Small Wind Turbines on Rooftops

A nice existing example of a wind turbine–building combination can be found at the Saxion Hogeschool

in Deventer (NL). [Figure 13](#) shows the building and the drag-driven VAWT on top of it.

The vertical axis wind turbine is used as an experimental setup for students. In the second instance, the similar wind turbine provides part of the electrical power for the illumination of the staircase. The wind turbines used on these buildings are modifications of existing products on the market.

Requirements for Wind Turbines on Rooftops The design of such urban wind turbines is driven by the following requirements:

- Good performance in complex winds
- Safe operation in the urban environment
- Low noise level
- Simple, rugged design
- Minimized maintenance
- Aesthetic appearance

With respect to the first point it already becomes clear that the design solution will look different for application on rooftops of buildings. HAWT-type wind turbines make use of a yawing system to track variations in wind direction. In urban areas, and definitively on top of buildings, wind direction variations can be considerable. HAWTs then have a clear disadvantage, since the yaw system is in general not able to follow these fast and extensive variations. This may



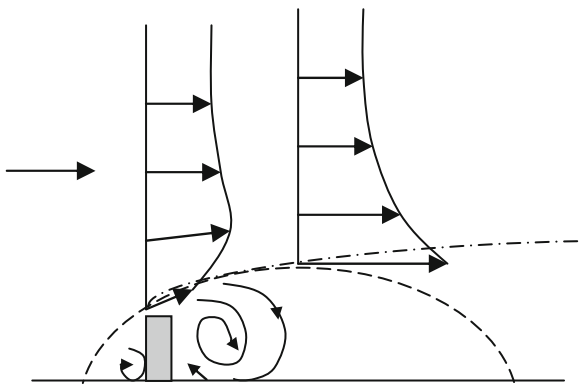
Electricity Generation with Small Wind Turbines. **Figure 13**

(a) Saxion building, Deventer, the Netherlands; (b) Drag-driven VAWT wind turbine on top of the central stairway

result in a continuous misalignment of the wind turbine with the direction of the wind, and an inherent loss in efficiency. This is a major reason for introducing VAWTs for rooftop application.

Flows over Rooftops Flows around buildings are inherently complex. At first, the topology of the terrain in the oncoming direction of the wind determines the local wind structure in terms of average wind speed; vertical gradient and turbulence intensity. Second, the topology of the building itself and of adjacent buildings will determine the local structure.

An example of the complex flow conditions that occur around a square building is shown in Fig. 14. Here the flow around a square obstacle is visualized. A separation bubble on top of the obstacle, starting at the leading edge can be observed as well as a closed cell vortex on the windward side and a large separation area with recirculation at the leeward side. Unfortunately, the details of the flow, such as depicted in Fig. 14, are very difficult to predict, even with state-of-the-art advanced computational (CFD) methods. Yet Mertens [15] showed that it is possible to develop a fairly simple theoretical model to determine the performance of wind turbines located in such complex flows. His calculations show that a significant flow acceleration effect is present on top of the building just outside the separating streamline. It is, however, restricted to a fairly small area. In this area, the wind will be about 20% higher than the undisturbed wind speed far in



Electricity Generation with Small Wind Turbines.

Figure 14

Flow over a high-rise building showing velocity increase and flow separation [14]

front of the building [14]. Figure 15 shows the result of the model for a high-rise building which is well exposed to the incoming wind.

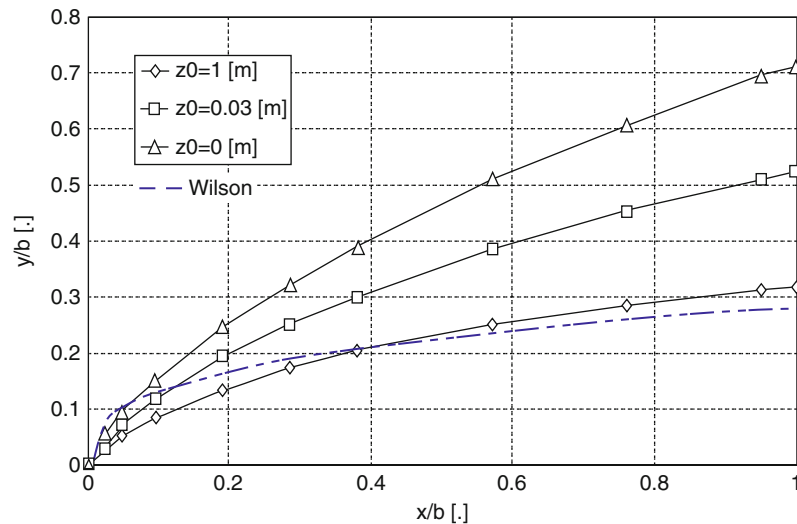
The results of the calculations of Mertens for a two-dimensional building (a building with infinite horizontal crosswind dimension) are shown in this figure for different values of the roughness height z_0 of the surface upwind of the building. The quantity b in this graph refers to the building depth, which is the horizontal size of the building in the direction of the incoming wind. Also shown is the result obtained by Wilson [21] for a square building. As can be seen there is a reasonable agreement between both results, for an upwind roughness height of 1 m, certainly for locations farther away from the windward edge of the building.

From further calculations of Mertens [15] it became clear that for a wind turbine located at the center of the roof a minimum height of 25% of the depth of a building is required for the lowest part of the rotor. For typical residential apartment buildings or office buildings this is in the order of 12–20 m. This means minimal tower height of 5 m underneath the rotor.

Suitable Wind Turbine Dimensions In the case the whole wind turbine is located in the higher wind speed region the energy increase is equal to the third power of 1.2. Thus, the aerodynamic power increase is 1.7 compared to a wind turbine in the same parallel flow without the building. When the wind turbine is located on the right spot on the roof this increase can be achieved for all wind directions. It, however, limits the dimensions of the wind turbine to something in the order of 20% of the characteristic height of a building. For uniform parallel oncoming flow over a mid- to high-rise its building height is equal to this characteristic height. Thus the limitation yields: $D \leq 0.2h$. For mid- to high-rise buildings with a characteristic dimension of say 50–100 m, the wind turbine diameter will, therefore, be from 10 up to about 20 m, in order to take full advantage of the speedup effect. Wind turbines of such sizes will generally not require significant modifications of the building structure, which is evidently an advantage when retrofitting buildings with renewable wind power.

Wind Turbines Developed for Operation on Buildings

A number of companies have seen the potential for wind turbine application on top of existing buildings



Electricity Generation with Small Wind Turbines. Figure 15

Location of the separating streamline for a two-dimensional building (20 m high, 10 m deep) [15]



Electricity Generation with Small Wind Turbines. Figure 16

High solidity and predominantly drag-driven VAWTs of various make

and have anticipated on the coming market with the development of dedicated machines. From the requirements for small wind turbines operating on buildings it is clear that vertical axis machines have a preference. Three examples of commercially available VAWTs for roof top application are shown in Fig. 16.

In similar fashion, examples of lift-driven VAWT designs are given in Fig. 17. It can be clearly seen that all

three designs are derivatives of the Darrieus wind turbine depicted in Fig. 2c. In the DWT design use is made of curved blades, just as in the original design. The Neoga® adds skewness of the blades, whereas the Turby® combines a cylindrical swept area together with blade sweep.

Since Turby® is the most deviating design, when compared to the original Darrieus design, it is worthwhile examining this particular design somewhat further.



Electricity Generation with Small Wind Turbines. Figure 17
Lift-driven VAWTs of various make

H-type Lift-Driven VAWT with Skewed Blades The three blades of the Turby® rotate along a cylindrical surface. Compared to the classical Darrieus rotor this has the advantage that a larger swept area is obtained for the same blade length. A disadvantage is the addition of struts in the center necessary for structural reasons. They will contribute to the viscous drag of the machine and thus reduce its efficiency. Furthermore, the stress path in the blades due to centrifugal loads is no longer aligned with the blade itself. At this point it should, however, be noted that the design “tip” speed ratio of Turby®, the ratio between circumferential speed and wind speed, is significantly less than for a classical Darrieus rotor (speed ratio $\lambda = 3$ for Turby® where $\lambda = 5\text{--}7$ for classical Darrieus rotors).

The blades are furthermore swept. This complicates the load path through the rotor further, but is applied because it has a very favorable effect upon noise production. The combination of low λ together with blade sweep makes it possible to comply with the noise levels allowable in the built environment. More details regarding the development and the design of Turby® can be found in Van Bussel et al. [5]. The Quit Revolution lift-driven VAWT is another implementation of the above described design philosophy (see Fig. 18).

The fact that these machines are not self-starting is often seen as a major disadvantage. However, when located on top of building the grid infrastructure is present to use the generator of the wind turbine temporarily used as a motor for kick-starting the machine at sufficient wind speeds.



Electricity Generation with Small Wind Turbines. Figure 18
Turby® prototype at TU Delft open-air research site

Standards and Certification

Generic design requirements for small wind turbines are provided in the standard IEC 61400–2, ed.2, “Wind Turbines – Part 2: Design Requirements of Small Wind Turbines” provided by the International Electrotechnical Committee.

Local standards and testing procedures for small wind turbines are developed by consensus through independent third-party authorities, such as the Institute of Electrical and Electronic Engineers (IEEE) and Underwriters Laboratories (UL) in the USA. Often use is made of the above mentioned IEC 61400–2 standard and of (components of) standards for large wind turbines such as IEC 61400-12-1 ed.1, “Wind Turbines – Part 12–1: Power Performance Measurements of Electricity Producing Wind Turbines” and IEC 61400–11 ed.2, “Wind Turbine Generator Systems – Part 11: Acoustic Noise Measurement Techniques,” and of other standards applied or modified for small wind turbines. According to AWEA [1], the standards UL 1741 and IEEE 929 have been successfully used to certify inverters for small wind turbines.

On top of that the IEEE is working on standard IEEE P1547 for intertied “distributed generation” technologies including small wind turbines [1].

Bibliography

1. AWEA (2009) AWEA small wind turbine performance and safety standard. AWEA 9.1 – 2009. American Wind Energy Association, Washington, DC
2. Bergey M Chair AWEA Small Wind turbine Committee (ed) (2002) Road map, a 20-year industry plan for small wind turbine technology. AWEA/Department of Energy, USA
3. van Bussel GJW (2003) Wind energy reliability data, a round-up from several data bases. In: Proceedings of the European wind energy conference & exhibition, Madrid, 16–19 June 2003
4. van Beveren SC (2008) Design of an urban wind turbine with diffuser. MSc thesis, TU Delft, Faculty of Aerospace Engineering, Delft University of Technology, Delft
5. van Bussel GJW, Mertens SM, Polinder H, Sidler HFA (2004) The development of turby, a small VAWT for the built environment. In: Proceedings of the global wind energy conference 2004, Chicago
6. van Bussel GJW (2007) The science of making more torque from wind: diffuser experiments and theory revisited. In: Proceedings of the science of making torque from wind. IOP, Lyngby
7. van Bussel GJW, Mertens SM (2005) Small wind turbines for the built environment. In: EACWE 4 conference proceedings, 11–15 July 2005, Prague, Paper No. 201
8. DWIA (2011) The guided tour. Danish Wind Industry Association, Denmark <http://wiki.windpower.org/>. Retrieved 20 Oct 2011
9. ETSU (1996) Working group on wind turbine noise, the assessment and rating of noise from wind farms. In: ETSU-R-97. ETSU, Harwell Didcot
10. IEC 61400–2, ed.2, Wind turbines – part 2: design requirements of small wind turbines. In: International electrotechnical committee, Geneva
11. Koehorst M (2001) De ontwikkeling van een windturbine voor toepassing in de gebouwde omgeving. M Sc thesis, TU Delft Faculty of Industrial Design, Delft
12. Lakeman LGJ (ed) (2002) Opwekking van windenergie in de gebouwde omgeving. Royal Haskoning, Nijmegen
13. Mertens SM (2001) Wind description for roof locating of wind turbines. Report 01173R, section Wind Energy. TU Delft, Delft
14. Mertens SM (2002) Notes on wind energy conversion in the built environment. Report 02187 R. TU Delft, Delft
15. Mertens SM (2005) Wind energy in the built environment. Phd thesis TU Delft, Faculty of Civil Engineering and Geosciences, Delft, The Netherlands
16. Put W (2000) Redesign of the Globuan Darrieus wind turbine. EJOK design for Industry, Rotterdam
17. Simiu E, Scanlan RH (1996) Wind effects on structures. Wiley, New York
18. Timmers G (ed) (2003) Eindverslag EET-Kiem project urban turbines. Ecofys, Utrecht
19. Toet S, Vermeer LJ (1996) Vermogensmetingen van een schaalmodel weerstandsturbine. Report IW96105R. Institute for Wind Energy, TU Delft, Delft
20. Van Dyke M (1968) An album of fluid motion. Parabolic, Stanford
21. Wilson DJ (1979) Flow patterns over flat-roofed buildings and application to exhaust stack design. ASHRAE Trans 28(2):284–294
22. Jacobs GJ (ed) (2003) Small wind turbine systems for battery charging. ARRAKIS, Veldhoven. www.arrakis.nl
23. de Jongh J, Smulders PT (2009) Introduction to wind pumps. ARRAKIS, Veldhoven. www.arrakis.nl
24. Lysen E (1983) Introduction to wind energy, CWD 82–1, 2nd edn. CWD, Amersfoort
25. Kentfield JAC (1996) The fundamentals of wind-driven water pumps. Overseas Publishing Association, Amsterdam. ISBN 2-88449-239-9
26. Renewable UK (2010) Generate your own power, your guide to installing a small wind system. Renewable UK, London
27. Tande OJ (1994) Estimation of cost of energy from wind energy conversion systems, IEA recommended practices for wind turbine testing. International Energy Agency. <http://www.ieawind.org/>
28. WinEur (2006) Wind energy integration in the urban environment, Techno Economic Report, Jan 2006
29. Wood D (2011) Small wind turbines, analysis, design and application. Springer, London. ISBN 978-1-84996-175-2

Electrochemical Hydrogen Production

MARC T. M. KOPER

Leiden Institute of Chemistry, Leiden University,
Leiden, The Netherlands

Article Outline

Definition of the Subject

Introduction

General Theory

Hydrogen Evolution at Metal Electrodes

Hydrogen Evolution at (Electrodes Modified with)
Molecular Catalysts in Aqueous and Nonaqueous
Solvents

Concluding Remarks

Acknowledgments

Bibliography

Definition of the Subject

The electrochemical production of hydrogen by photoelectrolysis or electrolysis is one of the most promising ways for making high-purity hydrogen. Water electrolysis is as old as the field of electrochemistry itself, and this entry focuses on some of the fundamental principles that have been established through the years.

Introduction

Hydrogen (H_2) is considered by many as the fuel of the future. In view of the current interest and importance in securing the sustainability of our society's energy needs, many research efforts are invested in making hydrogen, storing hydrogen, and converting hydrogen back to work. One of the most promising ways to make *clean* "solar hydrogen" is by electrochemistry or electrolysis, either in combination with a photovoltaic cell or in an integrated photo-electrochemical device. A crucial component of such a device is the catalyst converting water (or protons) into hydrogen. Such a catalyst must be efficient, stable, robust, cheap (or "cost-effective") and be made from abundant sources.

This entry will consider some fundamental considerations associated with the catalysis of the electrochemical hydrogen production. The entry will focus

on the energetic and thermodynamic requirements for such a catalyst [1–3] as opposed to the more usual approach (especially in the field of hydrogen evolution by molecular catalysts) in which emphasis is often on structural aspects of the catalyst. Of course, in an all-encompassing approach or theory, the two must be related, but in the author's view the approach emphasizing the energetic relationships provides a very useful general and basic framework for later, more detailed, approaches. The entry will also discuss these more detailed electrochemical models as well as the role of the solvent in the hydrogen evolution reaction (HER).

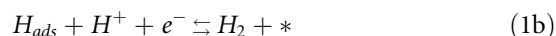
General Theory

Thermodynamic Theory

The hydrogen evolution reaction (HER), or the reverse hydrogen oxidation reaction (HOR), is a classic example of a two-electron transfer reaction



The reaction has a standard equilibrium potential E_{H_2/H^+}^0 of 0 V. The mechanism for the hydrogen oxidation/evolution reaction on metals will be considered first [4]. Later, this will be extended to a more general mechanism that encompasses the framework suggested for molecular catalysts. The classical electrochemical mechanism consists of three steps:



known as the Volmer, Heyrovsky, and Tafel steps, respectively. The symbol $*$ stands for a free adsorption site on the metal electrode surface, but it may also be considered as a coordination site for hydrogen in a molecular catalyst. The H_{ads} intermediate is a hydrogen species chemisorbed to such a site on the surface or in the molecular catalyst. Note that in reactions 1a and 1b, the proton and electron transfer are assumed to take place simultaneously. The concerted character of the proton–electron transfer for the Volmer reaction on Pt(111) is supported by recent detailed ab initio density functional theory (DFT) calculations [5, 6].

The above multistep mechanism needs to satisfy thermodynamic constraints in the sense that the combined equilibrium constants need to conform to the Nernst equation of the overall reaction [7]. Let us assume that the reaction follows a Volmer–Heyrovsky mechanism, in which case:

$$\frac{E_{\text{H}^+/\text{H}_{\text{ads}}}^0 + E_{\text{H}_{\text{ads}},\text{H}^+/\text{H}_2}^0}{2} = E_{\text{H}^+/\text{H}_2}^0 = 0 \quad (2)$$

The equilibrium potential of a redox reaction involving an adsorbed or chemisorbed state can be calculated if one defines the concomitant standard state. Consider the Volmer reaction 1a, for which the equilibrium condition is

$$\Delta G(\text{H}_{\text{ads}}) = \Delta G(\text{H}^+ + \text{e}^-) \quad (3)$$

Following Nørskov et al. [8], the term on the right-hand side is 0 at $E = 0$ (vs NHE) by definition, and $-e_0E$ at any other potential E , assuming that when the potential is changed only the energy of the electrons changes, and nothing else. The latter assumption may be tested by DFT calculations, showing that the binding energy of H_{ads} does not depend strongly on potential, or on an applied electric field [9–12]. This implies that the static dipole moment of the H_{ads} bond to the surface is close to 0, and that the electrosorption valency corresponding to reaction 1a is close to 1 [12]. Therefore, the standard equilibrium potential for the Volmer reaction follows as:

$$E_{\text{H}^+/\text{H}_{\text{ads}}}^0 = -\frac{\Delta G^0(\text{H}_{\text{ads}})}{e_0} \quad (4)$$

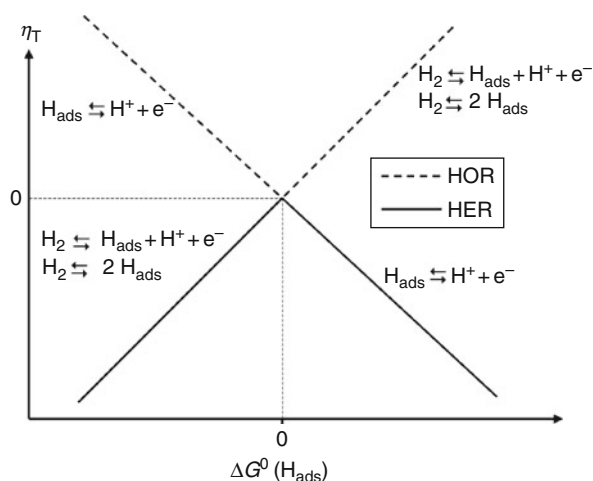
where $\Delta G^0(\text{H}_{\text{ads}})$ is the free adsorption energy of H_{ads} at some suitably defined standard state “0” (for instance, one-fourth coverage on a (111) surface), with respect to the energy of H_2 , which is equal to 0 also by virtue of the above definition. If $\Delta G^0(\text{H}_{\text{ads}}) < 0$, i.e., H_{ads} formation onto the catalyst surface yields a thermodynamically more favorable situation than H_2 evolution, $E_{\text{H}^+/\text{H}_{\text{ads}}}^0 > 0$. A well-known example of this is hydrogen underpotential deposition (UPD) on platinum and other transition-metal electrodes. Note, however, that in terms of the overall reaction, such a situation implies

$$E_{\text{H}^+/\text{H}_{\text{ads}}}^0 = -E_{\text{H}_{\text{ads}},\text{H}^+/\text{H}_2}^0 < 0 \quad (5)$$

so that the Heyrovsky reaction would have a standard equilibrium potential (defined for the same standard state) below 0 V. Neglecting kinetic factors, this analysis would suggest that if in the HER, H_{ads} formation would be a thermodynamically favorable reaction at $E = 0$, molecular hydrogen formation as in the Heyrovsky reaction must be thermodynamically unfavorable. Following Nørskov et al. [8], one could argue that in such a case, the minimal “thermodynamic overvoltage” for the overall reaction, η_{T} , would be equal to the standard equilibrium potential of the thermodynamically least favorable reaction, i.e., $\eta_{\text{T}} = E_{\text{H}^+,\text{H}_{\text{ads}}/\text{H}_2}^0 - E_{\text{H}^+/\text{H}_2}^0$. This “thermodynamically least favorable reaction step” is termed the *potential-determining step* (not to be confused with the rate-determining step, though often they may correspond to the same reaction step). Since the lowest possible “thermodynamic overvoltage” is 0, it follows that, on the basis of these thermodynamic considerations, we are looking for a catalyst with:

$$\eta_{\text{T}} = E_{\text{H}^+/\text{H}_{\text{ads}}}^0 = E_{\text{H}_{\text{ads}},\text{H}^+/\text{H}_2}^0 = \frac{\Delta G^0(\text{H}_{\text{ads}})}{e_0} = 0 \quad (6)$$

Any catalyst with $\Delta G^0(\text{H}_{\text{ads}}) \neq 0$ will have a $\eta_{\text{T}} > 0$ as either the Volmer reaction has a standard equilibrium potential more negative than 0 ($\Delta G^0(\text{H}_{\text{ads}}) > 0$) or the Heyrovsky reaction has a standard equilibrium potential more negative than 0 ($\Delta G^0(\text{H}_{\text{ads}}) < 0$). Figure 1 plots η_{T} for the HER as a function of $\Delta G^0(\text{H}_{\text{ads}})$, and illustrates how the optimal catalyst is a compromise between strong and weak binding of H_{ads} , as expressed by the classical Sabatier principle [13]. Figure 1 is similar but not identical to a volcano relationship, in which the catalytic activity for a reaction is plotted as a function of some property of the catalyst, in this case the catalyst’s ability to bind an intermediate. Typically such a plot leads to peak-shaped relationship of the catalytic activity versus the catalyst property, whence the name “volcano relationship.” Such a catalyst property is also known as a “descriptor,” as it gives a general but essential description of the catalyst properties. Although the “thermodynamic overpotential” of multistep electron transfer reactions is not a kinetic measure, the concept of the volcano relationship and associated Sabatier principle straightforwardly follow from thermodynamics combined with some very mild kinetic assumptions (one of the assumptions that one may make is that of a relation between activation energy



Electrochemical Hydrogen Production. Figure 1 “Thermodynamic” volcano plot for the hydrogen evolution reaction (HER) and the hydrogen oxidation reaction (HOR): thermodynamic overpotential (standard equilibrium potential of the potential-determining step) versus the free binding energy of hydrogen on the electrode surface [15, 17]. The dashed line gives the same plot for the HOR. The potential-determining steps corresponding to the various legs of the curves are indicated

and thermodynamic driving force, i.e., the Brønsted–Evans–Polanyi postulate [14]). Therefore, I consider Fig. 1 as a thermodynamic version of the kinetic volcano plot (i.e., $\log i_0$ vs $\Delta G^0(\text{H}_{\text{ads}})$, with i_0 the standard exchange current density), as originally suggested for the HER by Parsons [15] and Gerischer [16], and recently revived on the basis of DFT-calculated values for $\Delta G^0(\text{H}_{\text{ads}})$ by Nørskov et al. [17]. Note that the same analysis can be applied to the HOR, leading to the dashed-line volcano plot in Fig. 1, illustrating that for a reduction reaction we obtain a “^”-shaped plot, and for an oxidation reaction we obtain a “v”-shaped plot.

The same analysis may be applied to a Volmer–Tafel mechanism for the HER/HOR. The only difference now is that reaction 1c does not have a standard equilibrium potential, because it is not an electrochemical reaction. However, combined with reaction 1a, it must still satisfy the overall thermodynamic constraints of reaction 1. Defining the standard equilibrium constant of reaction 1c as:

$$K_{\text{H}_{\text{ads}}/\text{H}_2} = \exp\left[\frac{-\Delta G^0(\text{H}_{\text{ads}})}{RT}\right], \quad (7)$$

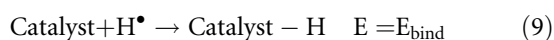
the Tafel–Volmer mechanism has to obey the following relation:

$$E_{\text{H}^+/\text{H}_{\text{ads}}}^0 - \frac{\Delta G^0(\text{H}_{\text{ads}})}{e_0} = 0 \quad (8)$$

In case $E_{\text{H}^+/\text{H}_{\text{ads}}}^0 < 0$, the thermodynamic overvoltage is $\eta_{\text{T}} = E_{\text{H}^+/\text{H}_{\text{ads}}}^0$, as in the Heyrovsky–Volmer mechanism. In case $E_{\text{H}^+/\text{H}_{\text{ads}}}^0 > 0$, the Tafel reaction becomes endothermic (Eq. 8), and therefore thermodynamically unfavorable. The most straightforward definition of a “thermodynamic overvoltage” would then follow from Eq. 8, leading to $\eta_{\text{T}} = -E_{\text{H}^+/\text{H}_{\text{ads}}}^0$. As a result, we see that the Heyrovsky–Volmer and the Tafel–Volmer mechanisms lead to “thermodynamic volcano curves” that are identical, and that more mechanistic assumptions are needed in order to distinguish between these two mechanisms. One important consequence of a volcano-type relationship such as that illustrated in Fig. 1 is that the potential-determining step is different on each side of the maximum. “Potential determining” means that this step determines the (over)potential at which a certain current density may be reached. For $\Delta G^0(\text{H}_{\text{ads}}) > 0$, the Volmer reaction is potential determining; for $\Delta G^0(\text{H}_{\text{ads}}) < 0$, either the Heyrovsky or Tafel reaction is potential determining. It is important to point out the difference between potential-determining and rate-determining step. Potential-determining steps are not always rate determining at the relevant potentials, as potential-determining steps suffer from an unfavorable thermodynamics, but not necessarily from an unfavorable kinetics at the potential at which the current finally starts flowing. Nevertheless, this simple analysis lucidly illustrates the thermodynamic constraints applicable to multistep multi-electron transfer reactions, and how catalysis may influence the “thermodynamic overvoltage” by optimizing the binding energy of the key intermediate, in this case H_{ads} . The best catalyst, according to this analysis, is the one that binds hydrogen in such a way that there is no thermodynamic sink or barrier in going from reactant through intermediate to product. This is the essential point made already by Parsons more than 50 years ago [15]. Experimental volcano plots of the exchange current density of the HER have been suggested many years ago by Trasatti, [18] and such plots may be criticized for various (experimental)

reasons [19, 20]. In this section, we will summarize some of the main objections. The section [Hydrogen Evolution at Metal Electrodes](#) will review the experimental evidence for the existence of a volcano relationship for the HER. At this point, it is stressed again that the approach discussed here does not include explicitly any kinetic aspects of the reaction.

The idea recently suggested by Nørskov et al. [17] is that $\Delta G^0(\text{H}^*)$ may be calculated for many potential catalysts from first principles by using DFT in combination with some mild assumptions on entropic contributions. In a DFT calculation, one typically calculates the binding energy of the hydrogen atom (or radical):



The free binding energy using H_2 as a reference can then be estimated from

$$\Delta G^0(\text{H}_{\text{ads}}) = E_{\text{bind}} - T^* \Delta S + \Delta \text{ZPE} + E(\text{H}^\bullet) \quad (10)$$

The term $T^* \Delta S$ takes into account the entropy contribution due to the vibration of hydrogen in its adsorbed state, and ΔZPE is the corresponding zero-point energy of the adsorbate. Their combined ($-T^* \Delta S + \Delta \text{ZPE}$) value has been estimated for H adsorbed on metals as ca. 0.24 eV by Rossmeisl et al. [11]. $E(\text{H}^\bullet)$ is half the bond energy of the hydrogen molecule, which is 2.24 eV (including the molecule's ZPE).

Trends in $\Delta G^0(\text{H}_{\text{ads}})$ or E_{bind} among different metals often conform to the so-called d-band model proposed by Hammer and Nørskov [14, 21]. This model claims a close correspondence between the strength of a surface bond (of H, but also of other adsorbates such as O, OH, and CO) and the location of the center of the d band of the metal surface. The energy of the d band may be “engineered” by changing the properties of the metal surface, for instance by (surface) alloying and overlayer formation [22]. The d-band center also influences the activation energy for dissociative adsorption, primarily through the Brønsted–Evans–Polanyi relationship [14], which assumes a linear relation between reaction energy and activation energy. The role of the d band in bond breaking electron transfer reactions was recently modeled by Santos and Schmickler [23], and will be discussed briefly in the next section.

The above mechanism, based on reactions 1a–1c, is rather specific to the HER on metal electrodes. If such a mechanism takes place on a molecular catalyst, the catalyst typically has two nearby (metal) centers to coordinate the hydrogen and adsorbed hydrogen is often referred to as “hydride.” The mechanism may be termed “homolytic” in this case, akin to the Volmer–Tafel mechanism. On many molecular catalysts or enzymes, however, the general mechanism tends to be different, and may be considered a variant of the Tafel–Heyrovsky mechanism, with the difference that the proton is not dissolved in the liquid phase, but also bound to a nearby “docking” site [1]. In such mechanisms, it is typically assumed that the release of the hydrogen molecule from the active site is an activated process (though the activation barrier may be small). Such a heterolytic mechanism may be written as follows:



where the two different coordination sites have been labeled as $*$ and $*'$. The two-electron transfer character of reaction 11 should be considered as formal; the charge distinction between H^- and H^+ may be exaggerated. Note that we have now the binding energies of H_{ch}^- and $\text{H}_{\text{ch}}^{+'}$ as the main descriptors for the thermodynamics of hydrogen evolution, so strictly speaking we have two descriptors now. It is not difficult to understand from Eqs. 11 that the properties of this mechanism will depend on the basicity/acidity of the coordinating sites $*$ and $*'$. As a result, the overall activity for hydrogen evolution will normally be strongly pH dependent. By contrast, hydrogen evolution on metals, which only involves neutral hydrogen as intermediate (at least formally), is typically not strongly pH dependent [1, 4].

More Detailed Theories and Simulations

The phenomenological theory of the previous section has an old history in the electrochemistry literature, and can be given a much more detailed treatment than

given here, for which reference is made to the original literature [15–18, 20]. More detailed molecular-level simulations of the hydrogen evolution (and hydrogen oxidation) reaction have been carried out in recent years, and it is of interest to discuss the main results from those investigations.

Santos and Schmickler consider that modeling hydrogen oxidation or hydrogen evolution fully from first principles is currently still too ambitious a task, and more general insight can be obtained from taking a simpler approach [23–26]. They have formulated a simple Hamiltonian which includes two hydrogens and the metal, and the electronic interaction between them [27]. The interaction with the solvent is treated in a spirit similar to the Marcus model. The ground-state energy of the model Hamiltonian may be calculated (with some approximations) and the interaction parameters entering the expression for the ground-state energy (strength of electronic interaction between hydrogen atoms; strength of electronic interaction of the hydrogen with the metal states; metal surface electronic band structure; strength of interaction of the hydrogen with the solvent) may be obtained from more detailed calculations, such as DFT. Although the effect of the simplifications of such an approach are somewhat difficult to estimate, the great advantage of this method is that it allows for a more detailed description of the HER/HOR without having to resort to extremely complicated calculations, which may have their own drawbacks. Moreover, the coupling between bond breaking and electron transfer can be followed in an instructive way. The activation energy in the theory of Santos and Schmickler is determined by the interaction of the bonding orbital of the hydrogen molecule with the d band of the metal catalyst as it passes the Fermi level of the metal. At this critical stage, the occupation probability of the bonding orbital is about one, so that one electron has already been transferred to the metal (in terms of the HOR). However, the activation energy for bond making or bond breaking does not only depend on the center of the d band as also predicted by the theory of Hammer and Nørskov. There is an additional important role of the electronic coupling between the hydrogen and the metal surface. The strength of the electronic coupling determines the broadening and splitting of the orbitals involved in the charge transfer reactions, and normally a stronger

electronic coupling leads to a lowering of the activation energy. Electronic coupling strengths typically increase toward the lower left-hand corner of the relevant part of the periodic table containing the coinage and transition metals, because of the diffuseness of the metal orbitals involved in the interaction increases in that direction. Based on their detailed calculations, Santos and Schmickler have criticized the use of volcano relationships in understanding trends in the metal-dependent reactivity of the HER/HOR. First of all, for most of the metals for which calculations can be performed, the data points lie on the ascending branch of the volcano curve. Metals that bind hydrogen very strongly also tend to bind oxygen very strongly, and therefore points on the descending branch of the volcano (increasing hydrogen binding strength) pertain to metals covered by an oxide film under aqueous electrochemical conditions. Hence, the experimental evidence for a volcano relationship is weak, if any, according to Santos and Schmickler. While this argument may be correct for the “older” volcano curves encompassing a wide variety of metals, it may not apply to some recent volcano curves (to be discussed in the section [Hydrogen Evolution at Metal Electrodes](#)). Secondly, the activation energy for hydrogen evolution and hydrogen oxidation depends not only on the strength of the interaction between hydrogen and the metal surface, but also on effects not (directly) incorporated into the simple model discussed in the previous section, such as the strength of the electronic coupling. The weak electronic coupling between nickel and hydrogen may explain the rather low activity of nickel, whereas the strong electronic coupling between rhenium and hydrogen may explain the high HER activity of rhenium [28]. Santos and Schmickler have argued that these two metals indeed appear to deviate from the volcano plot, precisely for the reason mentioned. A final point raised by Santos and Schmickler that is not explicitly included in the analysis of the previous section is that the binding energy of hydrogen is coverage dependent and that different metals may show their optimum activity at different coverages. As a result, Santos and Schmickler refer to the volcano curve, as applied to the HER/HOR as a “venerable” but effectively “defunct” and “oversimplistic” concept.

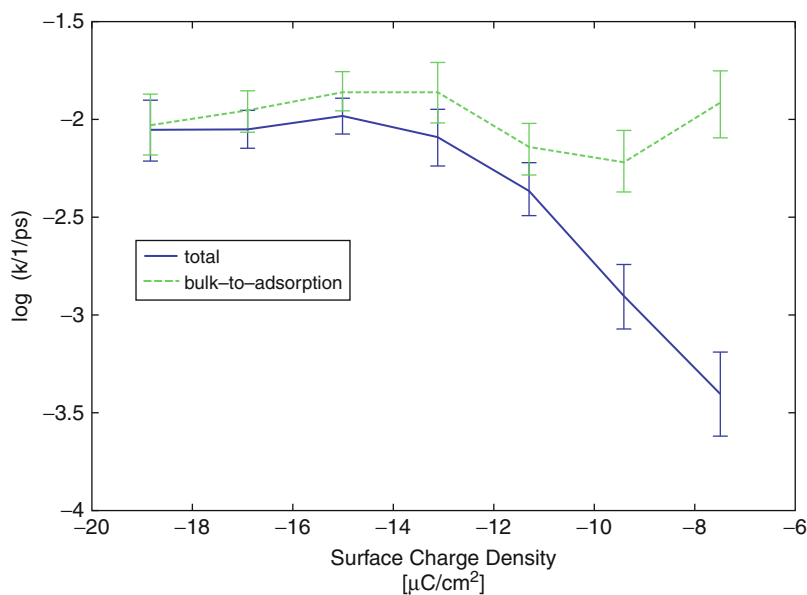
One of the simplifications made in the Santos–Schmickler approach to the HER/HOR is the way in

which the role of the solvent is included. The atomic level modeling of the solvent in the initial stages of proton reduction, without resorting to fully ab initio methods, was considered by Wilhelm et al. [29, 30]. They took the so-called empirical valence bond model (EVBM) to describe proton exchange in water to include proton transfer to a Pt(111) electrode surface. The model was parameterized on the basis of density functional theory calculations. The advantage of this approach over fully ab initio simulations (see below) is that much longer simulation times can be probed and therefore a more meaningful statistics can be obtained. This setup allowed the authors to study the dynamics of proton transfer to Pt(111) as a function of the charge density on the electrode surface. Figure 2 shows the charge dependence of two rate constants obtained from their detailed simulations: the inverse time needed for the proton to reach the “final” adsorption state, still being a proton, and the inverse time needed for “adsorption-to-discharge.” Figure 2 indicates that at strongly negative surface charge, the proton transfer reaction is dominated by the time the proton needs to approach the surface, and proton discharge is very fast. At less negative surface charge density, the rate for

establishing an orientational fluctuation suitable for proton transfer to the surface is rate determining.

Otani et al. [5] have carried out a full ab initio molecular dynamics simulation of the proton transfer reaction on a Pt(111) surface. This was the first ab initio dynamical simulation of an electrode reaction, but it should also be emphasized that this was a single-trajectory simulation, without statistical sampling, and that proton transfer was obtained only if excessively high electric fields were applied. Nevertheless, expect more of this kind of work in the future, and combined with rare-event sampling techniques, this is a potentially very powerful method. More recently, Santana et al. have used a similar approach to study the HOR on Pt(110) [31].

An alternative approach to studying the dynamic aspects of proton reduction and/or hydrogen evolution is to compute the potential energy surface for such reactions from detailed DFT calculations. A potential energy surface is generated at zero temperature, and therefore it does not include any finite temperature effects. Such an ab initio treatment of the hydrogen evolution reaction has been carried out recently by Skulason et al. [6, 32]. They included the



Electrochemical Hydrogen Production. Figure 2

Top: Logarithmic representation of the rate for proton discharge (blue, full line) and the proton bulk-to-adsorption rate (green, dashed) on Pt(111) (Reproduced with permission from Ref. [30])

effect of the electrode potential by considering various concentrations of protons (simulating different surface charge densities) within simulation cells of varying size, and extrapolating the obtained energies to infinite cell size to simulate a potentiostatic experiment. For the Volmer reaction on Pt(111), a barrier of ca. 0.5 eV was calculated. The Heyrovsky reaction was concluded to be the reaction with the highest barrier at 0 V, ca. 1.4 eV for the HER, whereas the Tafel reaction has a potential-independent barrier of ca. 0.8 eV at 0 V. As a result, the Tafel reaction is predicted to be the rate-determining step.

Hydrogen Evolution at Metal Electrodes

HER at Platinum Electrodes

As platinum is one of the best and most studied catalysts for the HER/HOR, as well as the one used in fuel cells and in many (acid) electrolyzers, it is useful to summarize here some of the pertinent experimental results. Platinum is generally agreed to be a very active catalyst for HER/HOR, and considered to be located near the apex of the volcano. Standard exchange current densities for the HOR are not easy to obtain because in general it is difficult to correct accurately for mass transport limitations and double-layer charging. Values from 10^{-3} to 1 A cm^{-2} have been reported for platinum [4, 33, 34] which would roughly correspond to a turnover rate of $0.1\text{--}100 \text{ s}^{-1}$ per Pt active site, at the standard equilibrium potential. The Pt “active site” is not well known, and there has been quite some controversy as to whether the HER/HOR is structure sensitive on platinum or not. In contrast to early studies, [35] the groups of Markovic [33] and Conway [34] have found that the Pt(110) surface is the most active in catalyzing the HER/HOR, typically being an order of magnitude more active than the other low-index surfaces.

In electrochemical experiments, rate-determining steps are usually inferred from so-called Tafel slopes, which essentially express the overall potential dependence of the reaction rate (i.e., the current) in a certain potential range. If the Volmer reaction is rate determining and its rate obeys the Butler–Volmer model, then the potential dependence is equivalent to a Tafel slope of $\ln 10 \times RT/(\alpha FE) \approx 120 \text{ mV/dec}$ if $\alpha \approx 0.5$. If,

on the other hand, the Heyrovsky reaction is rate determining, and the preceding Volmer reaction in “quasi-equilibrium,” the potential dependence of the reaction is equivalent to a Tafel slope of $\ln 10 \times RT/[(1 + \alpha)FE] \approx 40 \text{ mV/dec}$ if $\alpha \approx 0.5$. Finally, if the Tafel reaction is rate determining, the Tafel slope will be $\ln 10 \times RT/[2FE] \approx 30 \text{ mV/dec}$. In these predictions of the Tafel slope, saturation effects on the surface are not included.

Markovic et al. [33] have measured Tafel slopes for both the HER and the HOR on Pt single-crystal electrodes in acidic media. Tafel slopes also appear to depend on the crystal plane; the most active Pt(110) surface exhibits the 30 mV/dec Tafel slope well known from polycrystalline Pt [34, 36]. This Tafel slope corresponds to a Tafel–Volmer mechanism, i.e., a combination of reactions 1a and 1c.

The HER/HOR reaction kinetics is first order in the proton concentration over a wide pH range (0–14), implying that overpotentials are similar on the RHE (reversible hydrogen electrode) scale, in full agreement with the idea that proton and electron transfer are coupled over the entire pH range. However, there is a nontrivial pH effect in the sense that in alkaline media, the rate of the reaction tends to be about an order of magnitude lower than acidic media, at the same overpotential [37]. Conway et al. [37] have explained this effect by a more difficult abstraction of H from H_2O than from H_3O^+ . Pt(110) is still the most active surface, also in alkaline media, but the Tafel slope seems to be different, closer to 60 mV/dec , although no clear straight line was obtained.

An important discussion in the surface electrochemistry literature concerns the nature of surface-adsorbed hydrogen that acts as a precursor for hydrogen evolution in the HER, or for proton formation in the HOR [4, 38]. It is well known that at potentials above the reversible hydrogen potential, from ca. 0.05 to 0.3 V versus RHE (i.e., where no hydrogen evolution takes place), adsorbed hydrogen is formed on platinum. This form of adsorbed hydrogen is referred to as “underpotential deposited” (upd) hydrogen [38]. The surface coverage of this H_{upd} follows a Frumkin isotherm on Pt(111) and Pt(100) [38–40] with weak repulsive interactions between the adsorbed hydrogens. The coverages of H_{upd} in acidic and alkaline media on Pt(111) are almost identical at the same overpotential,

but again the kinetics of its formation is much slower in alkaline media [41]. The adsorbed hydrogen that participates in the HER at potentials around 0 V, is usually termed “overpotential deposited” (opd) hydrogen. The exact relation between H_{upd} and H_{opd} has led to ample discussion and speculation; for an illustrative account, see the paper by Conway and Jerkiewicz [42]. Most probably, H_{upd} and H_{opd} should primarily be distinguished on the basis of the different overall hydrogen coverages applicable to both states, with a typical H_{upd} coverage being lower than 0.5–1.0, and a typical H_{opd} coverage being close to or higher than 1 (i.e., including the H_{upd} coverage, implying more than 1 H per surface Pt atom), suggesting that H_{upd} and H_{opd} reside at different sites on the surface. The different coverages lead to different lateral interactions, different energetics, and site preferences, and therefore different activities toward H_2 evolution.

There have been a number of attempts to probe the adsorbed hydrogen on platinum using vibrational spectroscopy. A recent surface-enhanced infrared (IR) study on a platinum film electrode by Kunimatsu et al. [43] confirmed earlier results by absorption–reflection IR [44] that for potentials lower than ca. 0.1 V_{RHE} , atop-bonded Pt-H is formed. The signal of this band correlated with the hydrogen evolution current, and therefore this species was identified as the intermediate in the HER. Kunimatsu et al. suggest that H_{upd} adsorbs below 0.3 V_{RHE} in hollow sites, though this state has not been resolved spectroscopically, and that, in addition to H_{upd} , H_{opd} adsorbs below 0.1 V_{RHE} at atop sites, and serves as the reaction intermediate.

Experimental Evidence for Volcano Behavior on Strained Monolayer and Bimetallic Systems

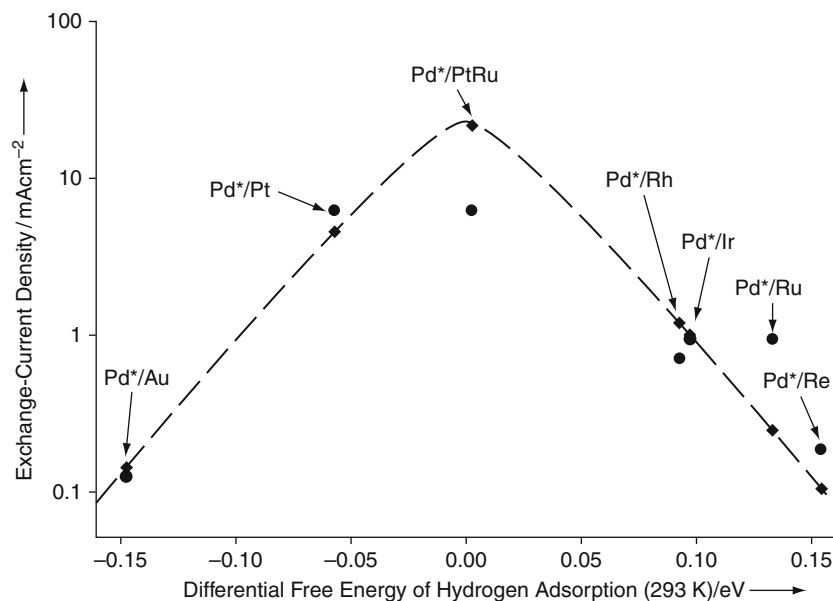
The evidence for a volcano-type relationship has been collected by Trasatti [18], and more recently by Nørskov and Stimming and coworkers [17]. One problem with such a correlation is that it is partially based on including metals that under experimental conditions are most likely covered by an oxide [20]. Recent studies by Kibler et al. [45, 46] have employed the idea of using palladium monolayers on different substrates [from Pd on Re(0001) to Pd on Au(111)] to tune the reactivity toward hydrogen adsorption and hydrogen evolution. Since the potential of the H_{upd} peak

observed in voltammetry is directly related to the adsorption energy of hydrogen (or more accurately, of H_{upd}), the authors could relate the peak potential to the theoretically predicted d-band shift with respect to Pd (111). The good correlation found suggested that this monolayer approach is a viable and controlled way to tune reactivity. In a subsequent study, the same systems were studied for their HER activity. Figure 3 shows the plot of the standard exchange current density versus the free energy of adsorption of hydrogen, clearly in agreement with the predictions of the Sabatier principle and the volcano plot. Note that from these results, an optimum standard exchange current density of ca. 0.01–0.1 $A\text{ cm}^{-2}$ is predicted.

Clearly, a catalyst consisting of Pd, Ru, and Pt is hardly an economical alternative for Pt, which is already quite a good catalyst for the HER/HOR. Therefore, Greeley et al. [47] have carried out an extensive computational study of bimetallic surfaces, screening for specific combinations with a $\Delta G^0(H^*)$ close to 0, with a negative formation energy to ensure stability of the material, and preferably consisting of cheaper elements. The counterintuitive combination of Pt and Bi emerged as one of the possibly attractive materials, with a $\Delta G^0(H^*)$ closer to zero than Pt. An appropriately prepared Pt-Bi surface alloy was subsequently found to indeed have a higher activity for the HER than Pt, showing the promise and usefulness of this computational screening strategy.

Hydrogen Evolution at (Electrodes Modified with) Molecular Catalysts in Aqueous and Nonaqueous Solvents

There have been ample efforts to develop molecular catalysts for hydrogen evolution, the majority of them inspired to mimic the activity of natural hydrogenases, resulting in an extensive literature on this subject that will not be summarized here. For some recent reviews, the reader may consult various recent sources [48–51]. These biomimetic complexes are usually organometallic compounds containing Fe, Ni, Co, and/or Ru. It is also known that certain nonmetal containing organic molecules (such as bipyridine) may enhance the HER on metal electrodes otherwise not particularly active for the HER, such as mercury or gold (see, e.g., Ref. [52] and references therein).



Electrochemical Hydrogen Production. Figure 3

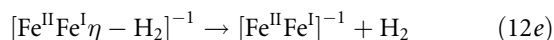
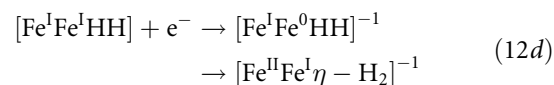
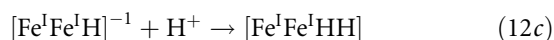
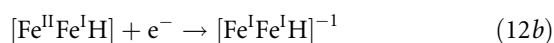
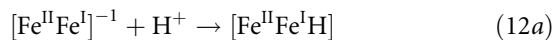
Hydrogen evolution exchange current density as a function of the theoretical differential free energy of hydrogen adsorption, $\Delta G^0(\text{H}^*)$. The *dashed line* shows the functional form of the theoretical exchange current density, and the *black diamonds* correspond to the predictions for specific overlayers. The *black circles* represent experimental measurements (the measured exchange current density determines the ordinate for these points, and the abscissa is specified by the theoretical $\Delta G^0(\text{H}^*)$ value). Pd*/X denotes a pure overlayer of Pd on a substrate of metal X (Reproduced with permission from Ref. [46])

Hydrogenases themselves have also been investigated as hydrogen evolution catalysts, and when adsorbed on graphitic electrodes, have been shown to exhibit very good catalytic properties, comparable to platinum. This latter subject has been discussed and reviewed extensively by Armstrong and coworkers [53]. To elucidate the catalytic function of these enzymes, it is useful to discuss some existing DFT calculations on computational hydrogenase models within the framework of the theory discussed in the section [General Theory](#).

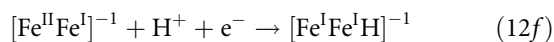
All hydrogenases catalyze both HER and HOR. The [FeFe] hydrogenases are most active for hydrogen formation, while [NiFe] hydrogenases are poised for hydrogen oxidation. Mechanistic information has been obtained from enzyme kinetics, spectroscopy, X-ray structural data, and DFT calculations. A general feature is that hydrogen binding or dissociation, proton exchange, and electron transfer are separate steps. A second distinct difference with the metal surface-catalyzed reaction is the heterolytic splitting or formation of hydrogen (reactions 11). A proton is abstracted

by a nearby base, and hydride is stabilized by the metal ions rather than H^* , although electron and spin densities are substantially delocalized. This is somewhat similar to the Heyrovsky reaction proposed for the mechanism on metals (reaction 1b). As mentioned, on a platinum electrode the homolytic bond splitting or making (Tafel reaction 1c) seems to be the preferred pathway.

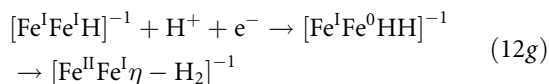
The mechanism of the HER/HOR by [FeFe] hydrogenases, as suggested by DFT calculations and experimental data, can be summarized as follows: [54]



The electrons are rapidly transferred from the reductant (or the electrode) to the active site via the chain of [FeS] clusters. From an electrochemical perspective, these [FeS] clusters thus are intramolecular mediators. Although Fe–H–Fe bridging intermediates have been proposed, the most likely mechanism involves one-Fe bound H-species and side-on bound H₂. The secondary amine in the bridging ligand and Fe^I appear to be equally potent as catalytic base. The iron most distant from the nearby [FeS] cluster, called the Fe_d site that shuttles between the formal 1+ and neutral state in reactions 12a–e, appears as the actual binding site for H and H₂. By taking together the proton and electron transfer steps in reactions 12a–d, we can compare the DFT calculations performed by Sbraccia et al. [54] to the model presented in the section General Theory. For reactions 12a and 12b together



the DFT-calculated reaction energy is −15.98 eV. Compared to the H₂ molecule, with a calculated total energy of 31.70 eV, the reaction energy is therefore −0.13 eV. For reactions 12c and 12d together:



the DFT-calculated reaction energy is −15.43 eV, implying a reaction energy of +0.42 eV. For the final H₂ release, reaction 12e, the reaction energy is −0.29 eV (with a calculated activation energy of 0.07 eV). Note that the three calculated reaction energies now add up to 0, implying that we are considering the reaction at thermoneutral conditions. Although this is an analysis based on total energies, as calculated by DFT, and not on free energies, these numbers still suggest that: (1) for hydrogenases, the key to an efficient catalyst is also to have an active site that binds H just as strong as H₂ (see also Ref. [55]), and (2) that the addition of the second hydrogen, reaction 12g, is thermodynamically uphill and hence the potential determining step, if we neglect zero-point energy contributions and entropy.

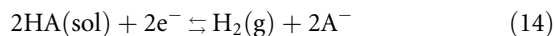
The organometallic nature of hydrogenase structural mimics makes them more soluble in nonaqueous solvents than in water. This has led to the development of activity tests in acetone, propylene carbonate, or acetonitrile for evaluation of model compounds. This

is obviously not without consequences for the definition of pH and of the equilibrium potential of the redox couples. The equilibrium potential of the reference couple, $E_{\text{H}_2/\text{H}^+}^0$, is indeed strongly dependent on the nature of the protic species (solvated protons or weak acid). The overpotential for hydrogen evolution, which is the relevant experimental quantity when comparing different catalysts in different solutions, must therefore be referred to the reversible potential of the H⁺/H₂ redox couple in the solvent under consideration. Working in a nonaqueous solvent has two important consequences for the determination of the reversible potential. First of all, the solvation energy of the proton is solvent dependent, and leads to a concomitant shift of the standard equilibrium potential. Second, many proton donors dissociate only incompletely in organic solvents, and therefore the dissociation constant should be taken into account in the definition of the standard equilibrium potential. Felton et al. [56] have described how to correct for both effects properly, and have given a useful comparison of various catalysts in a typical nonaqueous solvent, i.e., acetonitrile.

Considering the acid dissociation constant, $K_{a,\text{HA}}^{\text{sol}}$, in solvent “sol,” of the reaction:



the standard equilibrium potential of the overall half reaction



in the solvent “sol” is given by

$$E_{\text{H}_2/\text{HA}}^{0,\text{sol}} = E_{\text{H}_2/\text{H}^+}^{0,\text{sol}} - \left(\frac{2.303 RT}{F} \right) \text{p}K_{a,\text{HA}}^{\text{sol}} \quad (15)$$

In nonaqueous solvents, values for $E_{\text{H}_2/\text{H}^+}^{0,\text{sol}}$ or $E_{\text{H}_2/\text{HA}}^{0,\text{sol}}$ are often referred to the ferrocenium/ferrocene couple in the same solvent. The value of $E_{\text{H}_2/\text{H}^+}^{0,\text{aq}}$ would be a fundamentally more interesting reference, and if one is not interested in determining absolute electrode potentials, a simple way to look at the solvent dependence of $E_{\text{H}_2/\text{H}^+}^{0,\text{sol}}$, as referred to $E_{\text{H}_2/\text{H}^+}^{0,\text{aq}}$, is by considering the difference in the absolute solvation energies of the proton in “sol” and water:

$$E_{\text{H}_2/\text{H}^+}^{0,\text{sol}} = E_{\text{H}_2/\text{H}^+}^{0,\text{aq}} + \frac{[\mu^{\text{sol}}(\text{H}^+) - \mu^{\text{aq}}(\text{H}^+)]}{F} \quad (16)$$

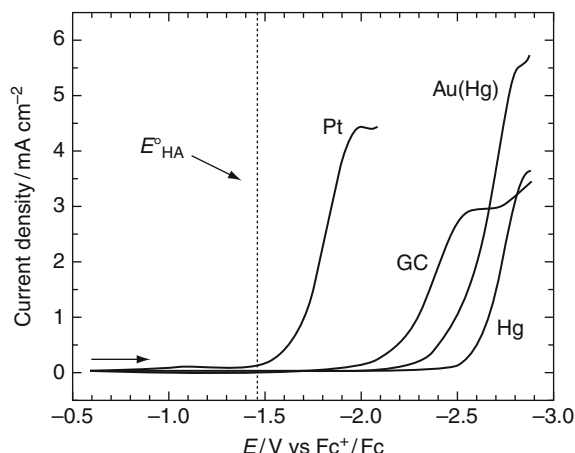
Values for the absolute solvation free energy (i.e., the Gibbs free energy of solvation not taking into account the surface potential of the solvent) of the proton in various solvents have been reconsidered recently by Fawcett [57] and the thus calculated values of $E_{\text{H}_2/\text{H}^+}^{0,\text{sol}}$ (vs $E_{\text{H}_2/\text{H}^+}^{0,\text{aq}}$) are given in Table 1. The experimental estimates of $E_{\text{H}_2/\text{H}^+}^{0,\text{sol}}$ have also been the subject of much work and debate. Fourmond et al. [58] have recently discussed their proper determination in a limited number of representative solvents, based on original data of Daniele et al. [59]. These values are, however, quoted versus an internal solvent standard, i.e., the ferrocenium/ferrocene Fc^{+0} couple. The equilibrium potential of this couple is usually considered solvent independent, although water appears to be an exception [60].

Using the tabulations provided in Ref. [60], the third and fourth columns in Table 1 have been generated, by first referring to decamethylferrocenium (DMFC)/decamethylferrocene couple (third column), which appears to be more solvent independent, and then rescaling such that in water the reference is at 0 V (fourth column). Although there is clearly a lack of data in this table to make any far-reaching conclusions, the differences between the theoretical and experimental values in acetonitrile and dimethylsulfoxide (DMSO) are typically within ± 0.15 V. More data are obviously needed to further elucidate this correspondence.

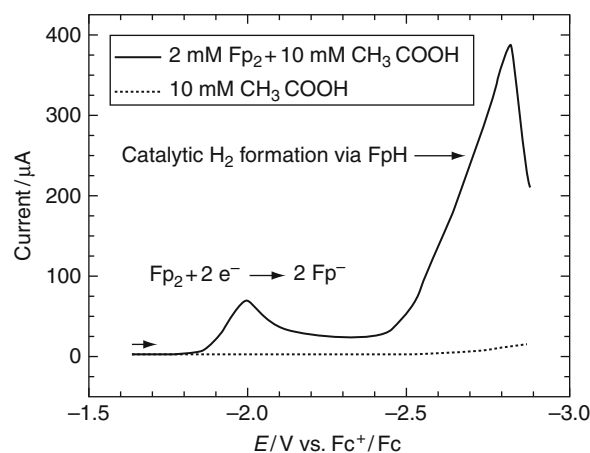
To illustrate the influence of a nonaqueous solvent on the HER activity, a comparison of different catalysts in acetonitrile is given in Figs. 4 and 5 [56]. It is seen

Electrochemical Hydrogen Production. Table 1
dependence of the equilibrium potential of the hydrogen electrode

Sol	$E_{\text{H}_2/\text{H}^+}^{0,\text{sol}}$ versus SHE/V (From Eq. 16, Ref. [57])	$E_{\text{H}_2/\text{H}^+}^{0,\text{sol}}$ versus $E_{\text{DMFC}^{+0}/0}$ /V (From Refs. [58, 60])	$E_{\text{H}_2/\text{H}^+}^{0,\text{sol}}$ versus SHE/V Combining [58, 60]
Water	0	−0.099	0
Acetonitrile	0.414	0.435	0.534
DMSO	−0.207	−0.202	−0.103
DMF	–	−0.162	−0.063
Methanol	0.083	–	–



Electrochemical Hydrogen Production. Figure 4
Linear sweep voltammograms of hydrogen evolution for 10 mM acetic acid in acetonitrile with 0.10 M tetrabutylammonium hexafluorophosphate at 1.0 V/s. Electrodes as indicated (GC is glassy carbon). Reversible potential for reduction of acetic acid shown for comparison (Reproduced with permission from Ref. [56])



Electrochemical Hydrogen Production. Figure 5
Linear sweep voltammograms for 2 mM Fp₂, an organometallic Fe compound, plus 10 mM acetic acid in acetonitrile with 0.10 M tetrabutylammonium hexafluorophosphate (solid curve) and 10 mM acetic acid alone (dashed curve). Hanging mercury drop electrode (area 0.0096 cm²). Scan rate: 1.0 V/s (Reproduced with permission from Ref. [56])

that Pt is still the best catalyst of the four catalysts illustrated in Fig. 4, and also significantly better than the hydrogenase mimic in Fig. 5, but the overpotential is rather high compared to the situation in an aqueous solvent. As a practical compromise to comparing overpotentials, it has been suggested to compare half-wave potentials. A detailed discussion on the proper comparison of these half-wave potentials has been given recently by Fourmond et al. [58].

Concluding Remarks

This entry has attempted to address some fundamental issues with studying the electrochemical production of hydrogen under various conditions and with various catalysts. The aim was to provide a common theoretical framework first, and then to show how to build on this framework and to understand the similarities and differences between the different catalytic approaches that have been used to tackle this important problem. It is hoped that this approach may be useful in ultimately developing a cheap, robust, and active catalyst for the electrochemical production of hydrogen. A main future direction for the development of new catalysts for electrochemical hydrogen production could be the combination of heterogeneous and molecular catalysis.

Acknowledgments

I am grateful to the Catalysis Research Center of Hokkaido University, Sapporo, Japan, for the visiting professorship during which this entry was written, and to Prof. Masatoshi Osawa for his kind hospitality. I am grateful to Dr. Vincent Fourmond for a useful discussion on the correct generation of Table 1.

Bibliography

Primary Literature

- Koper MTM, Heering HA (2010) In: Wieckowski A, Nørskov JK (eds) Fuel cell science: theory, fundamentals, and bio-catalysis. Wiley, New York, p 71
- Koper MTM, Bouwman E (2010) Angew Chem Int Ed 49:3723
- Koper MTM (2011) J Electroanal Chem (in press) doi: 10.1016/j.jelechem.2010.10.004
- Markovic NM, Ross PN Jr (2002) Surf Sci Rep 45:117
- Otani M, Hamada I, Sugino O, Morikawa Y, Okamoto Y, Ikeshoji T (2008) J Phys Soc Jpn 77:024802
- Skúlason E, Karlberg GS, Rossmeisl J, Bligaard T, Greeley J, Jonsson H, Nørskov JK (2007) Phys Chem Chem Phys 9:3241
- Bard AJ, Faulkner LR (2001) Electrochemical methods: fundamentals and applications. Wiley, New York
- Nørskov JK, Rossmeisl J, Logadottir A, Lundqvist L, Kitchin JR, Bligaard T, Jónsson H (2004) J Phys Chem B 108:17886
- Koper MTM, van Santen RA (1999) J Electroanal Chem 472:126
- Hamada I, Morikawa Y (2008) J Phys Chem C 112:10889
- Rossmeisl J, Nørskov JK, Taylor CD, Janik MJ, Neurock M (2006) J Phys Chem B 110:21833
- Koper MTM (2008) Faraday Disc 140:11
- Sabatier P (1911) Ber Deutsch Gem Ges 44:1984
- Nørskov JK, Bligaard T, Hvolbaek B, Abild-Petersen F, Chorkendorff I, Christensen CH (2008) Chem Soc Rev 37:2163
- Parsons R (1958) Trans Faraday Soc 54:1053
- Gerischer H (1958) Bull Soc Chim Belg 67:506
- Nørskov JK, Bligaard T, Logadottir A, Kitchin JR, Chen JG, Pandalov S, Stimming U (2005) J Electrochem Soc 152:J23
- Trasatti S (1977) J Electroanal Chem 39:163
- Petrii OA, Tsirlina GA (1994) Electrochim Acta 39:1739
- Schmickler W, Trasatti S (2006) J Electrochem Soc 153:L31
- Hammer B, Nørskov JK (1995) Nature 376:238
- Hammer B, Nørskov JK (2000) Adv Catal 45:71
- Santos E, Schmickler W (2007) Angew Chem Int Ed 46:8262
- Santos E, Schmickler W (2007) Chem Phys 332:39
- Santos E, Schmickler W (2008) Electrochim Acta 53:6149
- Santos E, Lundin A, Pötting K, Quaino P, Schmickler W (2009) Phys Rev B 79:235436
- Santos E, Koper MTM, Schmickler W (2008) Chem Phys 344:195
- Schmickler W, Quaino P, Santos E (2010) In: Abstract for the 61st ISE meeting, Nice
- Wilhelm F, Schmickler W, Nazmutdinov RR, Spohr E (2008) J Phys Chem C 112:10814
- Wilhelm F, Schmickler W, Spohr E, Phys J (2010) Condens Matter 22:175001
- Santana JA, Mateo JJ, Ishikawa Y (2010) J Phys Chem C 114:4995
- Skúlason E, Tripkovic V, Bjorketun M, Gudmundsdottir S, Karlberg G, Rossmeisl J, Bligaard T, Jonsson H, Nørskov JK (2010) J Phys Chem C 114:18182
- Markovic NM, Grgur BN, Ross PN (1997) J Phys Chem B 101:5405
- Barber J, Morin S, Conway BE (1998) J Electroanal Chem 446:125
- Gomez R, Fernandez-Vega A, Feliu JM, Aldaz A (1993) J Phys Chem 97:4769
- Guerrini E, Trasatti S (2009) Electrocatalysis in water electrolysis. In: Barbaro P, Bianchini C (eds) Catalysis for sustainable energy production. Wiley, Weinheim, p 235
- Barber JH, Conway BE (1999) J Electroanal Chem 461:80
- Jerkiewicz G (1998) Prog Surf Sci 57:137
- Koper MTM, Lukkien JJ (2000) J Electroanal Chem 485:161
- Garcia-Araez N, Lukkien JJ, Koper MTM, Feliu JM (2006) J Electroanal Chem 588:1
- Schouten KJP, van der Niet MJTC, Koper MTM (2010) Phys Chem Chem Phys 12:15217
- Conway BE, Jerkiewicz G (2000) Electrochim Acta 45:4075

43. Kunimatsu K, Senzaki T, Samjeské G, Tsushima M, Osawa M (2007) *Electrochim Acta* 52:5715
44. Nichols RJ, Bewick A (1988) *J Electroanal Chem* 243:445
45. Kibler LA, El-Aziz AM, Hoyer R, Kolb DM (2005) *Angew Chem Int Ed* 44:2080
46. Greeley J, Nørskov JK, Kibler LA, El-Aziz AM, Kolb DM (2006) *ChemPhysChem* 7:1032
47. Greeley J, Jaramillo T, Bonde J, Chorkendorff I, Nørskov JK (2006) *Nat Mater* 5:909
48. Artero V, Fontecave M (2005) *Coord Chem Rev* 249:1518
49. Radkowski Dubois M, Dubois DL (2009) *Acc Chem Res* 42:1974
50. Dempsey JL, Brunschwig BS, Winkler JR, Gray HB (2009) *Acc Chem Res* 42:1995
51. Artero V, Fontecave M (2010) *Energy Environ Sci* 3:727
52. Uchida T, Mogami H, Yamakata A, Sasaki Y, Osawa M (2008) *J Am Chem Soc* 130:10863
53. Armstrong FA, Belsey NA, Cracknell JA, Goldet G, Parkin A, Reisner E, Vincent KA, Wait AF (2009) *Chem Soc Rev* 38:36
54. Sbraccia C, Zipoli F, Car R, Cohen MH, Dismukes GC, Selloni A (2008) *J Phys Chem B* 112:13381
55. Hinneman B, Moses PG, Bonde J, Jørgensen KP, Nielsen JH, Hørch S, Chorkendorff I, Nørskov JK (2005) *J Am Chem Soc* 127:5308
56. Felton GAN, Glass RS, Lichtenberger DL, Evans DH (2006) *Inorg Chem* 45:9181
57. Fawcett WR (2008) *Langmuir* 24:9868
58. Fourmond V, Jacques P-A, Fontecave M, Artero V (2010) *Inorg Chem* 49:10338
59. Daniele S, Ugo P, Mazzochin G-A, Bontempelli G (1985) *Anal Chim Acta* 173:141
60. Noviadri I, Brown KN, Fleming DS, Gulyas PT, Lay PA, Masters AF, Phillips L (1999) *J Phys Chem B* 103:6713

Electrochemical Supercapacitors and Hybrid Systems

KATSUHIKO NAOI

Institute of Symbiotic Science and Technology, Tokyo
University of Agriculture & Technology, Koganei,
Tokyo, Japan

Article Outline

Glossary

Definition of the Subject

Introduction

Electrochemical Capacitor and Battery

Operation Voltage of Electrochemical Capacitors

Attempts to Increase Energy Density of Capacitors

Hybrid or Asymmetric Capacitor Systems

Lithium-ion Capacitor (LIC)
Nanohybrid Capacitor (NHC)
Future Perspectives
Bibliography

Glossary

Activated carbon Also called activated charcoal or activated coal is a form of carbon that has been processed to make it extremely porous and thus to have a very large surface area available for EDLC, adsorption, or chemical reactions.

Carbon nanotube Carbon nanotubes (CNTs), not to be confused with carbon fiber, are allotropes of carbon with a cylindrical nanostructure. Nanotubes have been constructed with length-to-diameter ratio of up to 132,000,000:1, significantly larger than any other material. These cylindrical carbon molecules have novel properties, making EDLC performances excellent in power capability.

Electrochemical capacitor (EDLC) An electric double-layer capacitor, also known as supercapacitor, pseudocapacitor, electric double layer capacitor (EDLC), supercapacitor or ultracapacitor is an electrochemical capacitor with relatively high energy density. Compared to conventional capacitors the energy density is typically on the order of thousands of times greater than an electrolytic capacitor. In comparison with conventional batteries or fuel cells, EDLCs have lower energy density but a much higher power density.

Energy density or Specific energy is defined as the energy per unit mass or volume.

Hybrid (asymmetric) capacitor A hybrid capacitor consists of a battery-like (faradic) electrode and a capacitor-like (non-faradic) electrode, producing higher working voltage and capacitance. With these systems, one can certainly achieve twice or triple enhancements in energy density compared to the conventional EDLCs.

Lithium-ion capacitor A Lithium-Ion Capacitor (LIC) is a hybrid type of capacitor. Activated carbon is used as cathode. The anode of the LIC consists of carbon material which is pre-doped with lithium ion. This pre-doping process lowers the potential of the anode and allows a high output voltage and higher energy density.

Nanohybrid capacitor A new lithium-ion-based hybrid capacitor using the lithium titanate ($\text{Li}_4\text{Ti}_5\text{O}_{12}$) negative intercalation electrode that can operate at unusually high current densities. The high-rate $\text{Li}_4\text{Ti}_5\text{O}_{12}$ negative electrode has a unique nano-structure consisting of extremely small nano-crystalline $\text{Li}_4\text{Ti}_5\text{O}_{12}$ nucleated and grafted onto carbon nano-fiber anchors (nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$ /CNF).

Power density Power density (or *volume power density* or *volume specific power*) is the amount of power (time rate of energy transfer) per unit volume.

Definition of the Subject

An electrochemical capacitor or electric double-layer capacitor (EDLC), also known as supercapacitor, pseudocapacitor, electrochemical double layer capacitor, or ultracapacitor is an energy storage device with high power and relatively high energy density. Compared to conventional electrolytic capacitors the energy density is typically 3 orders of magnitude greater. In comparison with conventional batteries or fuel cells, EDLCs show lower energy density but have a much higher power density. EDLCs have a variety of commercial applications, notably in energy-smoothing and momentary-load devices. They have applications as energy-storage devices used in vehicles, and applications like wind power solar energy systems where extremely fast charging required.

Historical Background

GE, General Electric engineers experimenting with devices using porous carbon electrodes first observed the EDLC effect in 1957. In 1966, SOHIO Standard Oil of Ohio developed the modern version of the devices, utilizing two layers of activated charcoal separated by a thin porous insulator. This basic cell design remains the basis of EDLCs. SOHIO failed to commercialize their device, licensing the technology to NEC, who finally marketed “supercapacitors” in 1978, as back-up of a computer memory. Around the mid-1990s various advances in materials science and refinement of the existing systems led to rapidly improving performance and an equally rapid reduction in cost. The first industrial applications were carried out for supporting the energy supply to robots.

In 2005, aerospace company, Diehl Luftfahrt Elektronik GmbH adopted EDLCs to power emergency actuation systems for doors and evacuation slides in airliners, including the new Airbus 380 jumbo jet. As of 2007 all solid state micrometer-scale electric double-layer capacitors based on advanced superionic conductors had been developed for low-voltage electronics.

Introduction

Environmental destruction are getting more and more serious in the global scale. It is urgent for human beings to cope with environmental and energy problems. The national development of each country greatly depends on economic activities oriented to environmental improvement business. Inevitably, in the future, it will be necessary to build an economic society taking into consideration the global ecology, that is, a society in which economic development is closely and directly related with eco-friendly activities such as carbon-reducing and energy-saving movements. Energy storage devices such as batteries and electrochemical capacitors hold an important key to solve the problems. The energy storage devices are some of the best fruits of our environmental technologies, and they have large potential for promoting our international contribution. Specifically, electrochemical capacitors are green devices having such excellent characteristics that they have long lifetimes and can be rapidly charged and discharged [7, 28, 31]. In particular, when combined with renewable energy plants (wind power, wave power and solar cell), they remarkably show their advantage and effectiveness [6]. In view of that, the electrochemical capacitors have been vigorously researched and developed in these days, and are expected to improve their energy densities. Thus, the technology is regarded as a promising art for storing electricity, and hence is expanding its possibilities.

Electrochemical Capacitor and Battery

Batteries store and deliver energy on the basis of chemical reactions, and hence their capacities depend on the redox reactions. In contrast, the energy storage or the charging–discharging mechanism by electrochemical capacitors are based on physical adsorption/desorption

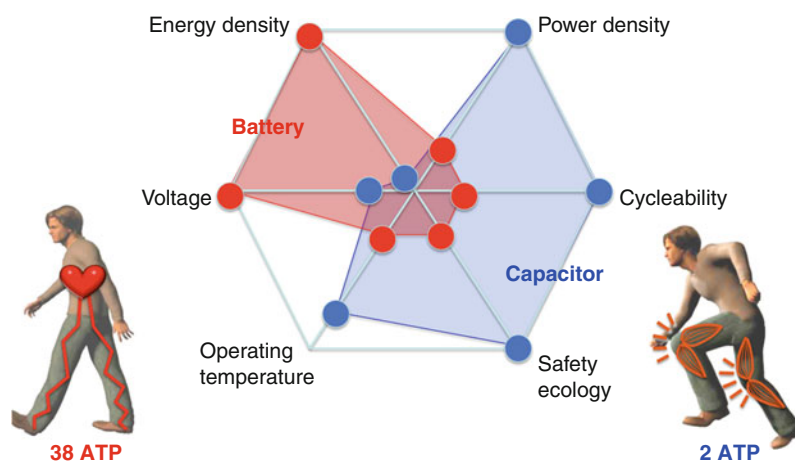
behaviors of ions at the interface of electrode/electrolyte, namely, electric double-layer [3, 6, 7, 28, 31]. Accordingly, the electrochemical capacitors are often called “electric double-layer capacitor (EDLC),” and more recently they are called “Supercapacitor” or “Ultracapacitor” as they increase their performances like six characteristics described below:

1. High power densities capable of instantly charging and discharging.
2. Long lifetimes capable of more than 100,000 charge–discharge cycles. Free from maintenance capable of using for a long time until abandoned.
3. Excellent safety and reliability issues. Low heat generation during operation. Recoverable even after short-circuiting.
4. Wide operational temperature ranges. More than 10°C lower and 20°C higher temperature limits as compared with any battery systems.
5. No fear of material shortage. Activated carbon materials can be obtained from any organic substance.
6. Free of heavy metals and halogens and Low environmental load.

As described above, the electrochemical capacitors have various advantages such as high power, high charge–discharge cycleabilities, high safety, and favorable working temperature ranges. The above characteristics are remarkably contrastive with those of lithium

ion batteries as shown in Fig. 1. The other cannot replace one of these devices. The contrasting two types of energy devices as represented by lithium ion batteries and electrochemical capacitors can be compared to two different energy-providing systems in a human body, i.e., aerobic reaction system and anaerobic reaction system. It is very interesting to see a human body from the viewpoint of its energy system. In a human body, ATP (adenosine triphosphate) efficiently takes charge of energy production and supply. The ATP is mainly produced by way of two routes (i.e., aerobic reactions and anaerobic reactions). The aerobic reactions, which proceed with oxygen obtained by breathing, slowly produce no less than 38 ATP molecules (from one glucose molecule), to provide large energy. On the other hand, the anaerobic reactions can rapidly produce two ATP molecules, to provide small but instant energy according to necessity. The combination of these systems realizes such smooth energy supply as can correspond to any movement of human body. In the same way, the combinations of lithium ion batteries and capacitors give very high synergy effects. Consequently, energy management can be conducted under any situation by use of the two contrasting energy devices in exquisite combination.

In many uses, lithium ion batteries and capacitors play complementary roles. Since the synergy effects obtained by their combinations are very high, practical



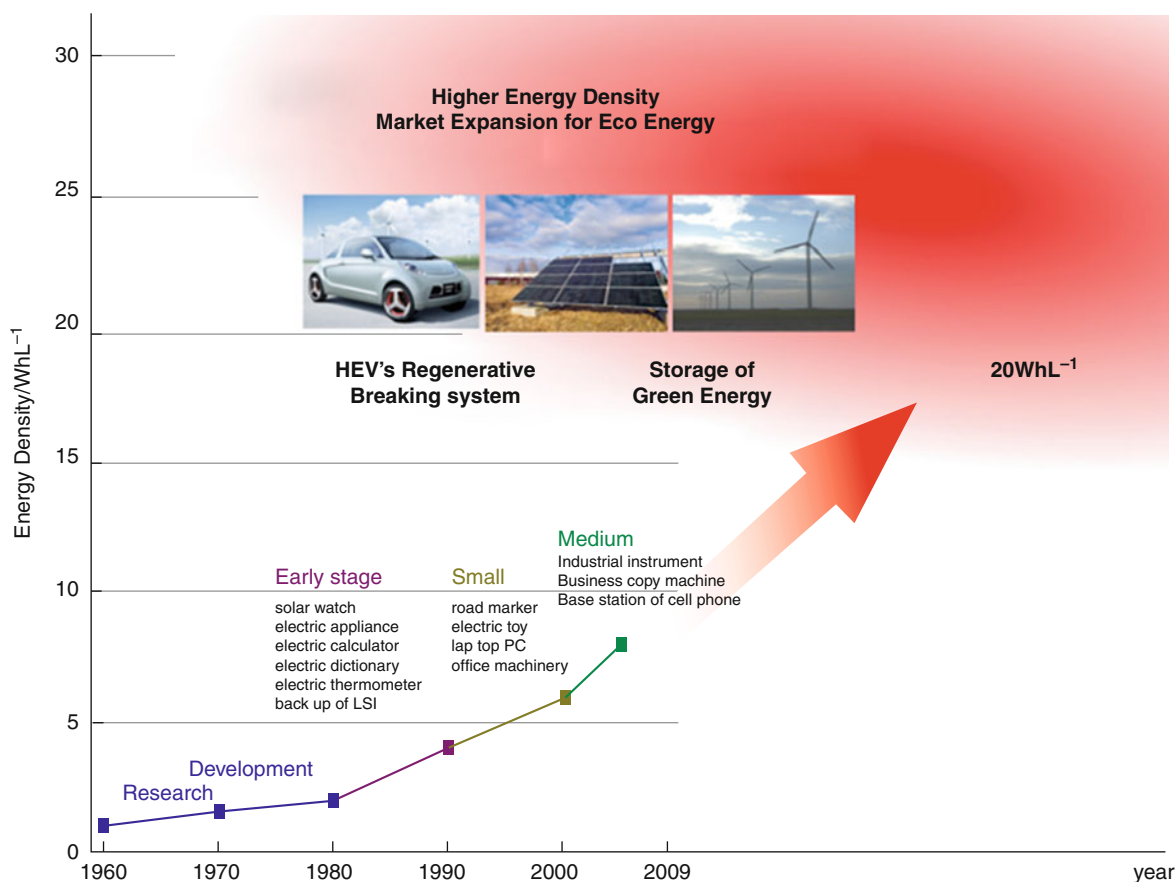
Electrochemical Supercapacitors and Hybrid Systems. Figure 1
Beautiful contrast of battery and capacitor

examples thereof are increasing in various fields. Further, the capacitors may be combined with devices other than the lithium ion batteries, for example, with fuel batteries or motor/inverters. At present, researchers pay their attention to smart systems employing the combinations of the capacitors and those devices. In consideration of necessity for effective and eco-friendly energy measures, it will be one of the large strategic aims to develop the composite energy storage systems.

Operation Voltage of Electrochemical Capacitors

However, since the EDLCs generally have low energy densities, their uses are limited and they cannot fully meet various performance demands required by the recent markets as shown in Fig. 2. Particularly in the

field of automobiles, new energy devices are strongly desired to have hybrid characteristics between lithium ion batteries and EDLCs and thereby which can be suitably employed in idle reduction systems. Accordingly, it is expected for them to form a large market [28]. In order to satisfy the performance demands, it is often said to be necessary for the EDLCs to enhance their energy density to $20\text{--}30\text{ Wh L}^{-1}$, which is approximately twice or more than present EDLCs, namely, $5\text{--}10\text{ Wh L}^{-1}$. For realizing the above high energy density, hybrid capacitor systems comprising non-aqueous redox materials are being dynamically researched and developed in recent years [5, 6, 10, 16, 23, 31]. The present chapter deals with the recent contributions to get this high energy density by focusing on two major hybridized cell configurations in organic media.



Electrochemical Supercapacitors and Hybrid Systems. Figure 2

Target value of the next-generation electrochemical capacitor

As is described above, increasing energy density is one of the most crucial matters. For conventional EDLC systems, designed with two symmetrical activated carbon electrodes, increasing voltage is more effective because the energy density increases in proportion to the squared voltage. Thus, it is essential to develop higher electrochemical durability at the electrode/electrolyte interface. Currently, the maximum voltage of an activated carbon-based EDLC (AC/AC) is limited to 2.5–2.7 V in Propylene carbonate (PC) and Acetonitrile (AN) electrolytes. At applied voltage over 2.7 V, there occurs a significant decrease in capacitance and an increase in internal resistance. In fact the float test revealed that the AC/AC capacitor lose its capacitance ($\Delta C < -13\%$ after 30 days) at 2.5 V, whereas at 2.9 V the capacitance loss became more significant ($\Delta C = -28\%$) for the same duration. The threshold cell voltages between 2.5 and 2.9 V certainly triggers a consecutive and fatal degradation. The undesired faradic process (failure modes) that leads to a capacitance fade is the most critical factor that determines the life of conventional EDLCs and needs to be further investigated as in the recent review article by Simon et al. [28]. Anyway this withstanding-voltage limitation (2.5–2.7 V) certainly exists because of a gas evolution at higher voltages and is a barrier for further enhancement of the energy density of the conventional activated-carbon-based EDLCs.

Attempts to Increase Energy Density of Capacitors

There is presently a major effort to increase the energy density of EDLCs up to a target value in the vicinity of 20–30 Wh kg⁻¹ [7]. Many studies have been undertaken in order to enhance the energy density (E) of EDLC. There are mainly three approaches (Fig. 3): the first is to change the electrode (by higher-capacitance carbons or redox) and the second is to change electrolyte (by durable new electrolyte or ionic liquids).

① High working voltage(V)

② High double layer capacitance(C)

$$E = 1/2 (CV^2)$$

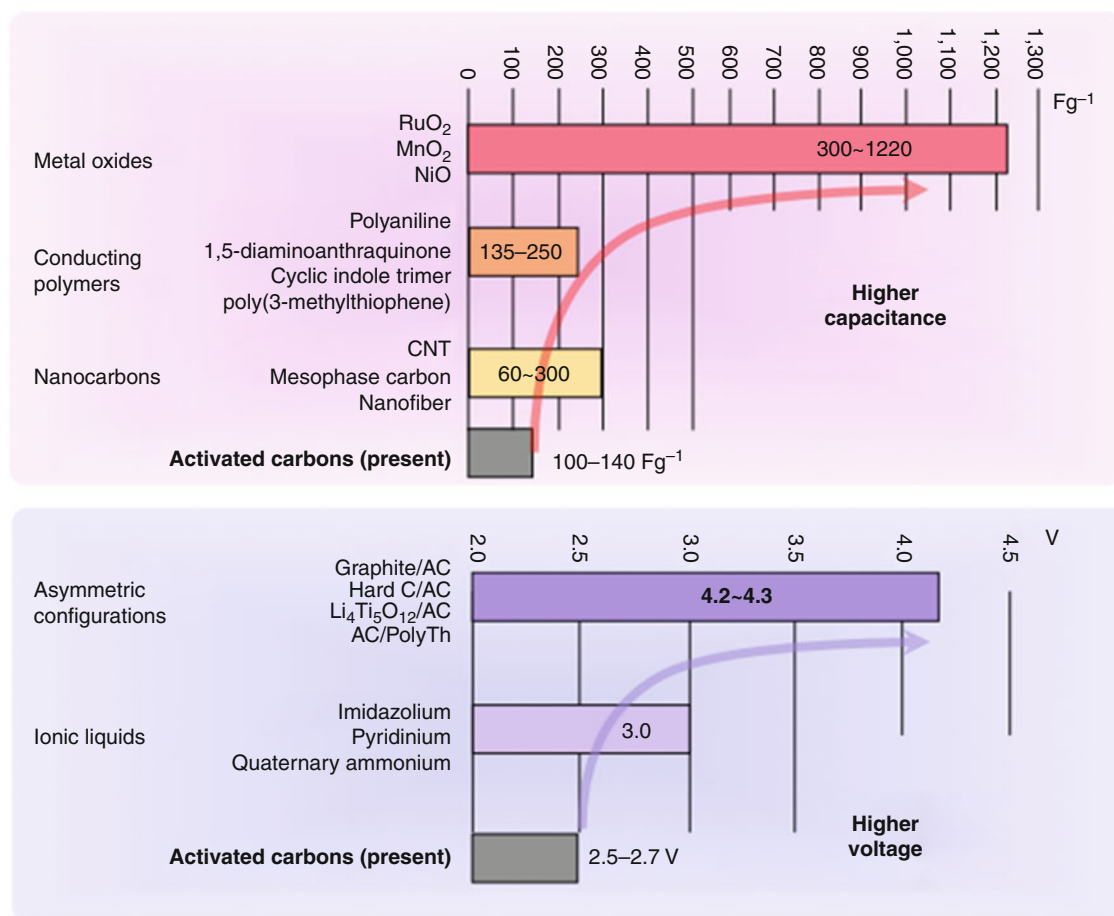
$$\left(\begin{array}{l} c : \text{Capacitance (F)} \\ v : \text{Working voltage(V)} \end{array} \right)$$

Hybrid or Asymmetric Capacitor Systems

The third and the most important approach to meet the goal that is under serious investigation is to develop asymmetric (hybrid) capacitors. As shown in Fig. 3 various hybrid capacitor systems are possible by coupling redox-active materials (e.g., graphite [10, 31], metal oxides [5, 14], conducting polymers [17, 18]), and an activated carbon (AC). There are suggested two different systems, one is aqueous and the other is non-aqueous. Some of them are listed below in Fig. 4. These approaches can overcome the energy density limitation of the conventional EDLCs because it employs a hybrid system of a battery-like (faradic) electrode and a capacitor-like (non-faradic) electrode, producing higher working voltage and capacitance. With these systems, one can certainly achieve twice or triple enhancements in energy density compared to the conventional EDLCs.

FDK corporation proposes a “Dual Carbon Cell,” which is a device behaving “like a capacitor” on the basis of faradic reactions on the electrodes. When the Dual Carbon Cell is charged, anions and cations (such as lithium ions) are intercalated into the positive graphite electrode and into the negative graphite electrode, respectively. As a result, an acceptor type graphite intercalation compound is formed on the positive electrode while a donor type one is formed on the negative electrode. Thus, the Dual Carbon Cell is charged or discharged like a capacitor is. It is reported that, if the positive electrode is made of graphite doped with boron, the Dual Carbon Cell can have improved load characteristics.

Meanwhile, many other hybrid capacitors are also proposed each of which utilizes a charge transfer reaction (faradic reaction) on either positive or negative electrode whereby one could have a good scenario attaining higher voltage and higher energy density. Figure 5 represents the voltage profiles of the conventional EDLC system (upper), Nanogate (middle), and LIC (lower). Examples of them include a “nanogate” capacitor or a “nanostorage” capacitor (Fig. 5, middle), in either of which the negative electrode is an activated carbon but the positive electrode is an amorphous carbon capable of undergoing an anion-intercalation



Electrochemical Supercapacitors and Hybrid Systems. Figure 3
Energy density enhancement of electrochemical capacitors

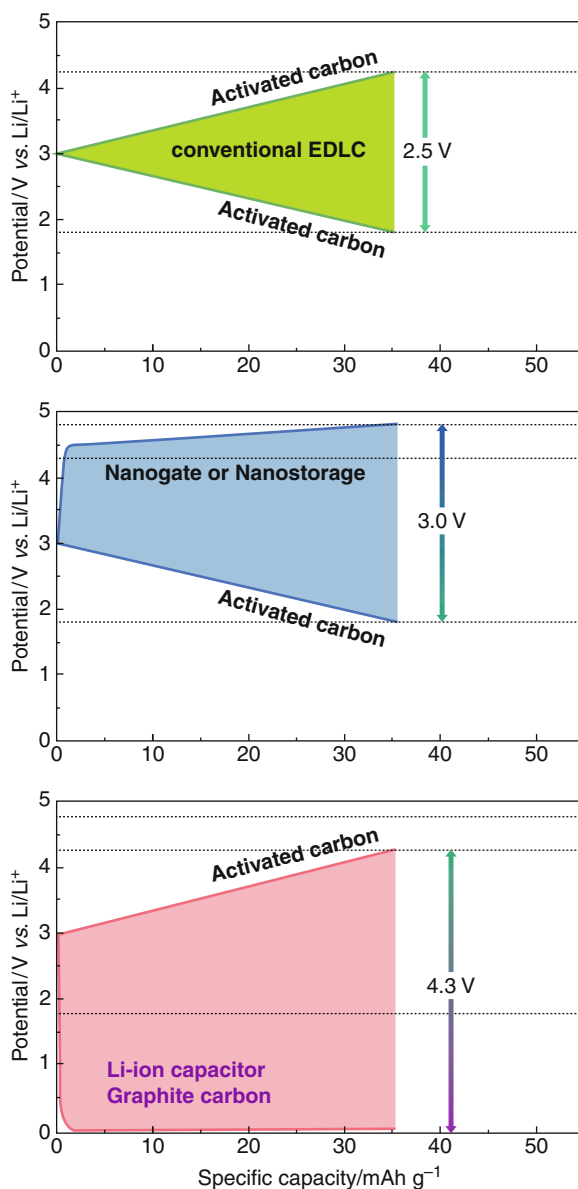
Aqueous hybrid capacitor	Non-aqueous hybrid capacitor	
Carbon/redox	Carbon/Carbon	Carbon/redox
AC/LiMn ₂ O ₄ AC/PbO ₂ AC/NiO	Graphite/Graphite Graphite/AC Graphite(Li ⁺)/AC AC/Graphite	Li ₄ Ti ₅ O ₁₂ /AC Li ₄ Ti ₅ O ₁₂ /AC+LiCoO ₂ AC/pMeTh Li ₄ Ti ₅ O ₁₂ /pPPhTh

Electrochemical Supercapacitors and Hybrid Systems. Figure 4
Aqueous and non-aqueous hybrid capacitor systems

reaction at a high potential (not lower than 4.5 V vs. Li/Li⁺), while a normal capacitor comprises both positive and negative electrodes made of activated carbon (Fig. 5, upper).

Lithium-Ion Capacitor (LIC)

Among high-energy hybrid capacitors comprising non-aqueous redox materials, a hybrid system called “lithium-ion capacitor (abbreviated as LIC)” has particularly attracted the attention in these days [10, 31]. The LIC is a hybrid capacitor in which the positive and negative electrodes are made of activated



Electrochemical Supercapacitors and Hybrid Systems.

Figure 5

Voltage profiles of a conventional EDLC, Nanogate, and LIC systems

carbon and of graphite pre-doped with lithium ions, respectively (Fig. 6).

Since the negative electrode of graphite undergoes the reaction at a potential a little over 0 V vs. Li/Li^+ , the LIC has a high working voltage of 3.8 V to 4.0 V. This high working voltage enables the lithium-ion

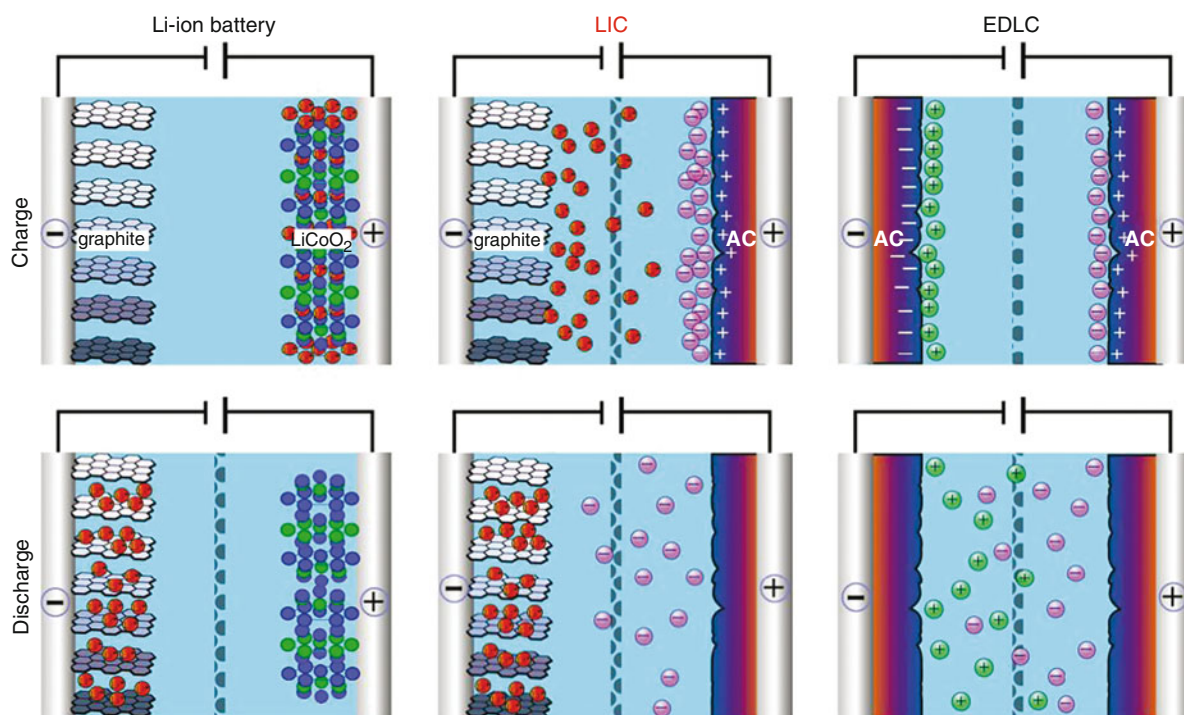
capacitor to realize both a high power density of approximately 5 kW kg^{-1} and an energy density of approximately $20\text{--}30 \text{ Wh kg}^{-1}$. The LIC thus exhibits favorable performance, and is regarded as a candidate of the next generation electrochemical capacitor. Hence, some Japanese companies including JM energy, FDK, ACT, etc. have already started commercializing LIC [10].

Generally speaking, instead of achieving energy density improvement, the high working voltage causes an electrolyte decomposition problem especially at negative graphite electrode [21] (Fig. 7). The limited charging rate and the low temperature performance are the possible drawbacks of the LIC because it has a lithium-intercalation negative electrode. The process of pre-lithiation of the graphite electrode may lead to poor cost-effectiveness or instability in the quality of the LIC device when they are mass-produced (see Fig. 8). This “pre-doing of Li^+ ” may bring about the same risks relevant to a long-term stability to keep low ESR. This is an issue that is specifically important for an electrochemical capacitor as a “power device” because it lead to create a high impedance electrode/electrolyte interface and thus eventually leads to a deteriorated power performance for longer cycles.

Nanohybrid Capacitor (NHC)

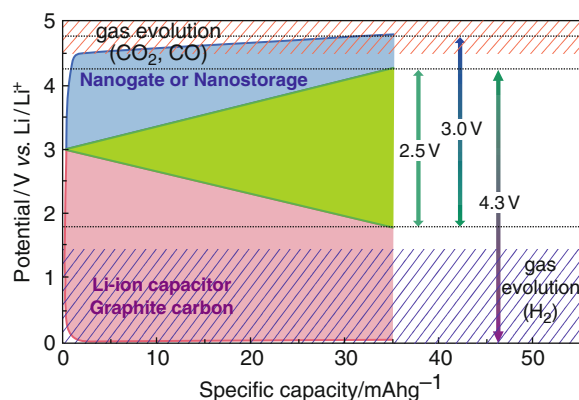
Very recently, Naoi's group and Nippon Chemi-con developed a hybrid capacitor system that certainly achieves a high energy density, high stability, and high safety. This is called “Nanohybrid Capacitor” (abbreviated as NHC) using a super-high-rate nano-structured lithium titanate ($\text{Li}_4\text{Ti}_5\text{O}_{12}$)/carbon composite negative electrode. They have kept their eyes on lithium titanate ($\text{Li}_4\text{Ti}_5\text{O}_{12}$) as a stable and safe redox material capable of increasing the energy densities of hybrid capacitors without sacrificing interfacial characteristics. The $\text{Li}_4\text{Ti}_5\text{O}_{12}$ operates at a potential (1.55 V vs Li/Li^+) out of the range where the electrolyte solution may be decomposed, and hence play key roles for providing capacitor systems as stable and safe as EDLCs (see Fig. 9). The $\text{Li}_4\text{Ti}_5\text{O}_{12}$ as a redox material for hybrid capacitors has the following essential advantages in energy density, in stability, and in safety:

1. It exhibits a high coulombic efficiency close to 100% during charge–discharge cycle



Electrochemical Supercapacitors and Hybrid Systems. Figure 6

Cell configurations of Li-ion Battery (LIB), EDLC, and LIC



Electrochemical Supercapacitors and Hybrid Systems. Figure 7

Figure 7

Reliability and electrolyte decomposition at positive and negative electrodes of LIC systems

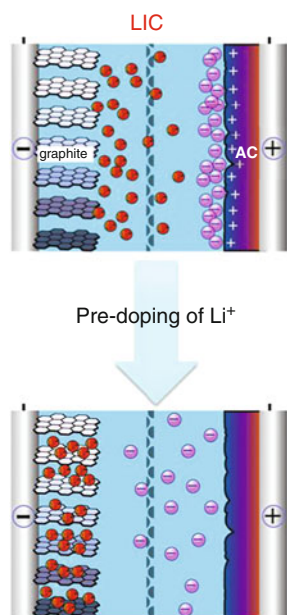
2. It has a theoretical capacity (175 mAh g^{-1}) four times higher than as activated carbon
3. It undergoes charging and discharging at a constant potential of $1.55 \text{ V vs. Li/Li}^+$, where the electrolyte

solution is free from the fear of decomposition (no SEI formation and no gas evolution) [20, 27]

4. As shown in Fig. 10, it changes the volume to a very small degree (0.2%) in charging or discharging in intercalating or deintercalating Li^+ ions (zero-strain insertion)
5. Inexpensive raw material

As summarized in Table 1, a capacitor cell comprising $\text{Li}_4\text{Ti}_5\text{O}_{12}$ is, therefore, expected to be free from necessity for pre-doping of lithium ions (because of the above advantage (1)), to be capable of having a high energy density (because of the advantage (2)), and to be excellent in stability and in safety (because of the advantages (3) and (4)).

Amatucci et al. firstly introduced the $\text{Li}_4\text{Ti}_5\text{O}_{12}/\text{AC}$ system as safer hybrid systems [1, 2]. However, the conventional $\text{Li}_4\text{Ti}_5\text{O}_{12}$ has the greatest problem of low power characteristics that stem from inherent poor Li^+ diffusion coefficient ($<10^{-6} \text{ cm}^2 \text{ s}^{-1}$) [29] and poor electronic conductivity ($<10^{-13} \text{ } \Omega^{-1} \text{ cm}^{-1}$) [8].

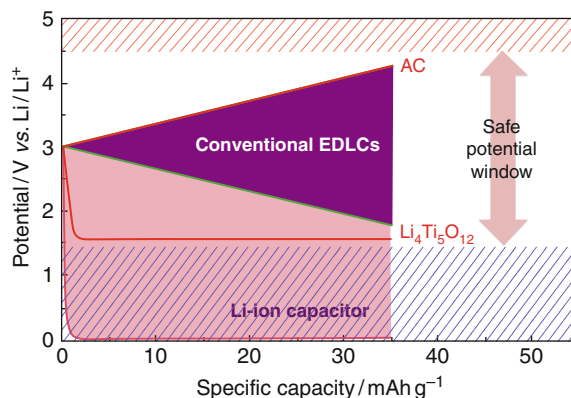


Electrochemical Supercapacitors and Hybrid Systems.
Figure 8

Necessary pre-doping of Li^+ for LIC systems

Accordingly $\text{Li}_4\text{Ti}_5\text{O}_{12}$ had a big problem of slow output characteristics and was not fully developed for the application in electrochemical capacitors at that time. For the purpose of solving the problem of poor output performance, some measures can be thought of. For example, the $\text{Li}_4\text{Ti}_5\text{O}_{12}$ may be ground down to nano-sized particles and may be combined with an electro-conductive material to prepare a composite.

In the nano-sized $\text{Li}_4\text{Ti}_5\text{O}_{12}$ particles, it is expected that Li^+ ions diffuse and electrons migrate in distances reduced by 1/1,000 (see Figs. 11 and 12) or less as compared with the distances in normal micron-order $\text{Li}_4\text{Ti}_5\text{O}_{12}$ particles. Further, in the composite, electron paths are effectively formed in the electrode of nano-sized $\text{Li}_4\text{Ti}_5\text{O}_{12}$ particles (Fig. 13). On the basis of these expectations, a composite (nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$ /CNF) has been synthesized by means of a new method referred to as “Ultracentrifugal force (UC) nanohybridization method” [19, 24]. Specifically, the composite (nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$ /CNF) comprises nano-crystalline



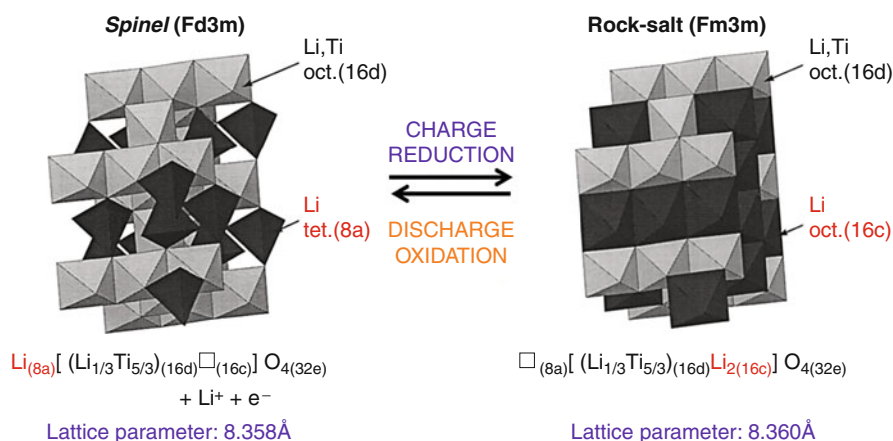
Electrochemical Supercapacitors and Hybrid Systems.

Figure 9

Voltage profile of conventional EDLC, LIC, and NHC

$\text{Li}_4\text{Ti}_5\text{O}_{12}$ particles hyper-dispersed on carbon nano-fibers (CNFs) [13] having high electro-conductivity ($25 \Omega^{-1}\cdot\text{cm}^{-1}$). They utilized the composite as the negative electrode active material, and thereby succeeded in producing a novel hybrid capacitor (nanohybrid capacitor) realizing both high power and high energy density.

The nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$ negative electrode was developed to have a unique nano-structure that can operate at unusually high current densities. Nano-crystalline $\text{Li}_4\text{Ti}_5\text{O}_{12}$ attached onto carbon nano-fibers were prepared by a unique technique (UC method) of a mechano-chemical sol-gel reaction under ultracentrifugal force field [13, 19, 24], followed by an instantaneous heat treatment for very short duration (see Fig. 14a). The UC method is induced to form, anchor, and graft the nano-Li-Ti-O precursors on the CNF matrices. The subsequent instantaneous heat treatment is of prime importance to precisely achieve all of the following: high crystallization of $\text{Li}_4\text{Ti}_5\text{O}_{12}$, [13] inhibition of the CNF oxidation decomposition during the annealing at high temperature, [31] and suppression of the agglomeration of the $\text{Li}_4\text{Ti}_5\text{O}_{12}$ particles. These processes are quite simple and require only a few minutes. Actually, the power characteristic of the prepared composite (nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$ /CNF) made a new bench mark (charge-discharge time = 12 s)



Electrochemical Supercapacitors and Hybrid Systems. Figure 10

Structural and volumetric changes at the charged and discharged states of $\text{Li}_4\text{Ti}_5\text{O}_{12}$

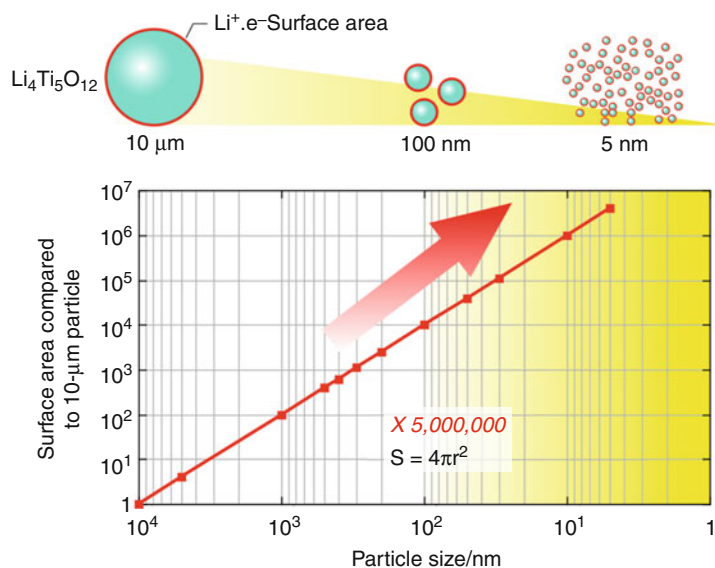
Electrochemical Supercapacitors and Hybrid Systems. Table 1 Comparison of LIC and NHC in their cell configurations

LIC		VS.	NHC	
negative	positive		negative	positive
AC Graphite/HC/PAC Li pre-doping necessary Li-salt / EC, PC Higher than EDLC Al (pos), Cu hole (neg) -10°C			AC, CNT nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$ /CB, CNF, CNT No Li pre-doping Li-salt / PC, DMC, AN EDLC comparable Al (pos & neg) -40°C	
Manufac. procedure Electrolyte Internal resistance Current collector Low-temp. perform.			Positive electrode Negative electrode Manufac. procedure Electrolyte Internal resistance Current collector Low-temp. perform.	

[4, 11, 12, 15, 24, 32] which exceeds greatly the shortest charge–discharge time that has ever been attained (see Fig. 14b).

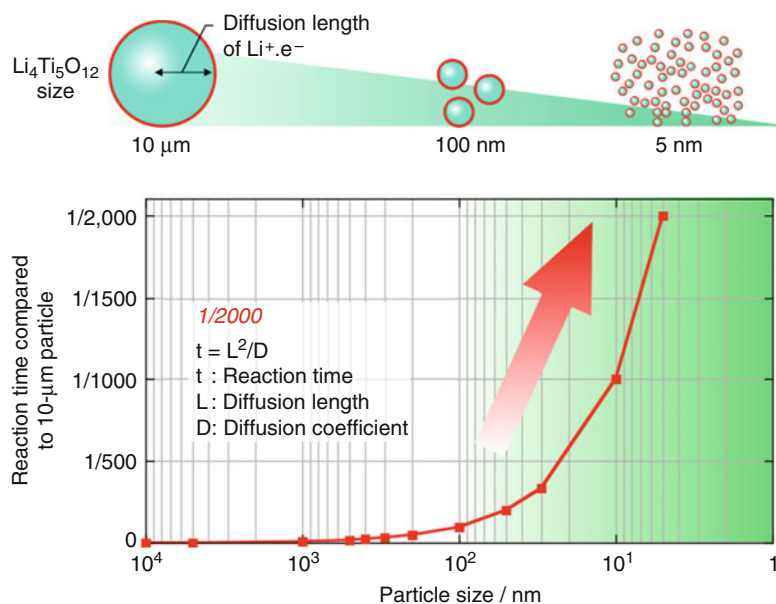
XRD analysis was performed to confirm the formation of the nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$ and the presence of the CNF in

the nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$ /CNF composites. Figure 15a shows the XRD patterns of the prepared nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$ /CNF composite and pristine CNF. The composite has several sharp diffraction peaks at $2\theta = 18, 35, 42, 57,$ and 63° . These peaks correspond to (111), (311), (400), (511),



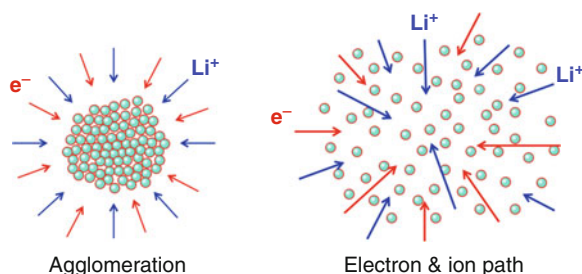
Electrochemical Supercapacitors and Hybrid Systems. Figure 11

Increased accessibility and surface area as a function of the particle size of $\text{Li}_4\text{Ti}_5\text{O}_{12}$



Electrochemical Supercapacitors and Hybrid Systems. Figure 12

Decreased diffusion length as a function of the particle size of $\text{Li}_4\text{Ti}_5\text{O}_{12}$



Electrochemical Supercapacitors and Hybrid Systems.

Figure 13

Hyper-dispersion of the nano-size particles of $\text{Li}_4\text{Ti}_5\text{O}_{12}$

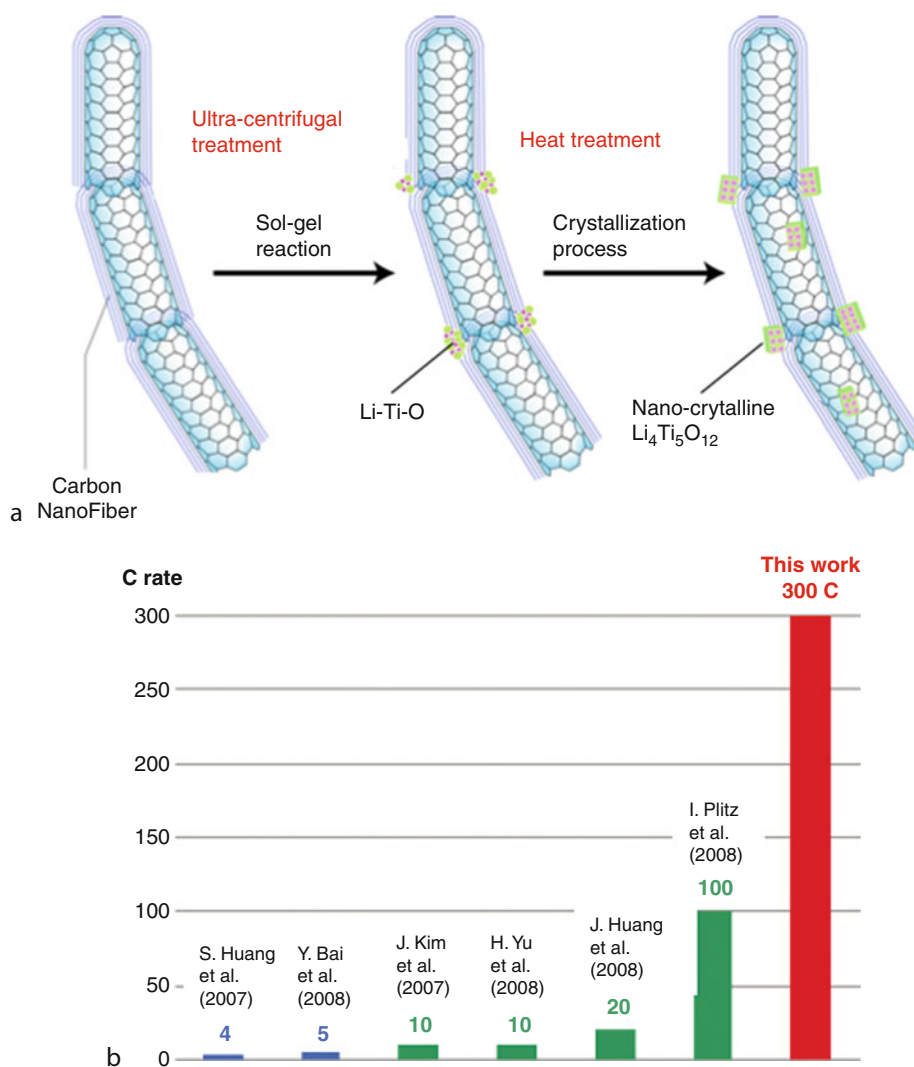
and (440) planes of a face-centered cubic spinel structure with $Fd\bar{3}m$ space group [22, 30], respectively, indicative of the formation of crystalline $\text{Li}_4\text{Ti}_5\text{O}_{12}$. Annealing at 900°C under vacuum well crystallizes the nano-particles of $\text{Li}_4\text{Ti}_5\text{O}_{12}$. A broad peak at around $2\theta = 24.5^\circ$ is observed which corresponds to the (002) plane of the pristine CNF [33]. This means that the CNF exists in the annealed composite and preserves its graphene layer structure. The fact that there are no other peaks observed corresponding to some possible impurities such as TiO_2 , Li_2CO_3 , and Li_2TiO_3 [9, 26] suggests that there are only two species, crystalline $\text{Li}_4\text{Ti}_5\text{O}_{12}$ and CNF.

Thermogravimetric (TG) measurement of the nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$ /CNF composite was performed under air to estimate the residual weight ratio of the CNF. The obtained TG curve is shown in Fig. 15b. The weight loss at $400\text{--}600^\circ\text{C}$ resulted by the oxidative decomposition of the CNF, and exactly 50 wt.% of the nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$ remained. The 50 wt.% is well consistent with the weight ratio of the $\text{Li}_4\text{Ti}_5\text{O}_{12}$ to the CNF calculated on the basis of dosed Ti alkoxide weight before UC method. This fact implies that sol-gel reaction in the UC method stoichiometrically proceeds, and the optimized (very short duration) annealing does not cause oxidative decomposition of the CNF. Such a stoichiometric preparation process (UC method and instantaneous annealing) is one of the important factors for cost-effectiveness of capacitor production.

The nano-structures and crystallinity of nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$ and CNF in the composite is observed by HR-TEM (see Fig. 16). The image indicates that the edge or defect sites of CNF graphenes accommodate and graft nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$ particles. The results support that the $\text{Li}_4\text{Ti}_5\text{O}_{12}$ particles are formed through a nucleation process onto the edge and defect CNF sites. The clear facet of $\text{Li}_4\text{Ti}_5\text{O}_{12}$ reflects the high crystallinity that is consistent with the sharp XRD spectrum (see Fig. 16a), despite such a nanosize. Such a high crystallinity resulted in a reversible, smooth Li^+ insertion performance, and ca. 100% of coulombic efficiency. Also, HR-TEM images of CNF show clear graphene layers indicative of crystalline structure. Thus, this composite is considered to be the junction material of two crystalline species, $\text{Li}_4\text{Ti}_5\text{O}_{12}$ and CNF. Of particular interest is that the lattice matching of the nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$ particles, and CNFs are perfect and they are firmly tied at the atomic level. This could bring about an establishment of good electronic paths between the two species (Fig. 16b).

Another closer view of an HR-TEM (see Figure 17) indicates the multiple attachments of the nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$ particles on the inside wall of the CNF graphene substrate as well as outside wall of the CNF matrices. This may indicate that the crystallization would be such efficient on CNF and more importantly one can enhance both the energy density and the specific gravity of the nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$.

Figure 18a shows the charge-discharge curve of the half-cell $\text{Li}/(\text{nc-}\text{Li}_4\text{Ti}_5\text{O}_{12}/\text{CNF})$ at 1 C. The horizontal axis represents specific capacity per unit weight of $\text{Li}_4\text{Ti}_5\text{O}_{12}$. The dominant plateau was observed at ca. 1.5 V versus Li/Li^+ that corresponds to the Li^+ intercalation-deintercalation process of the crystalline $\text{Li}_4\text{Ti}_5\text{O}_{12}$ [13, 24, 30], indicating that the capacity of the composite is determined by the redox capacity of nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$ in the composite. The obtained capacity was 167 mAh g^{-1} per $\text{Li}_4\text{Ti}_5\text{O}_{12}$, which is 95% of the theoretical capacity. It is noted here that the value of 167 mAh g^{-1} is obtained after subtracting the double-layer capacity of the CNF (8 mAh g^{-1}). This result indicates that almost all of the nc- $\text{Li}_4\text{Ti}_5\text{O}_{12}$



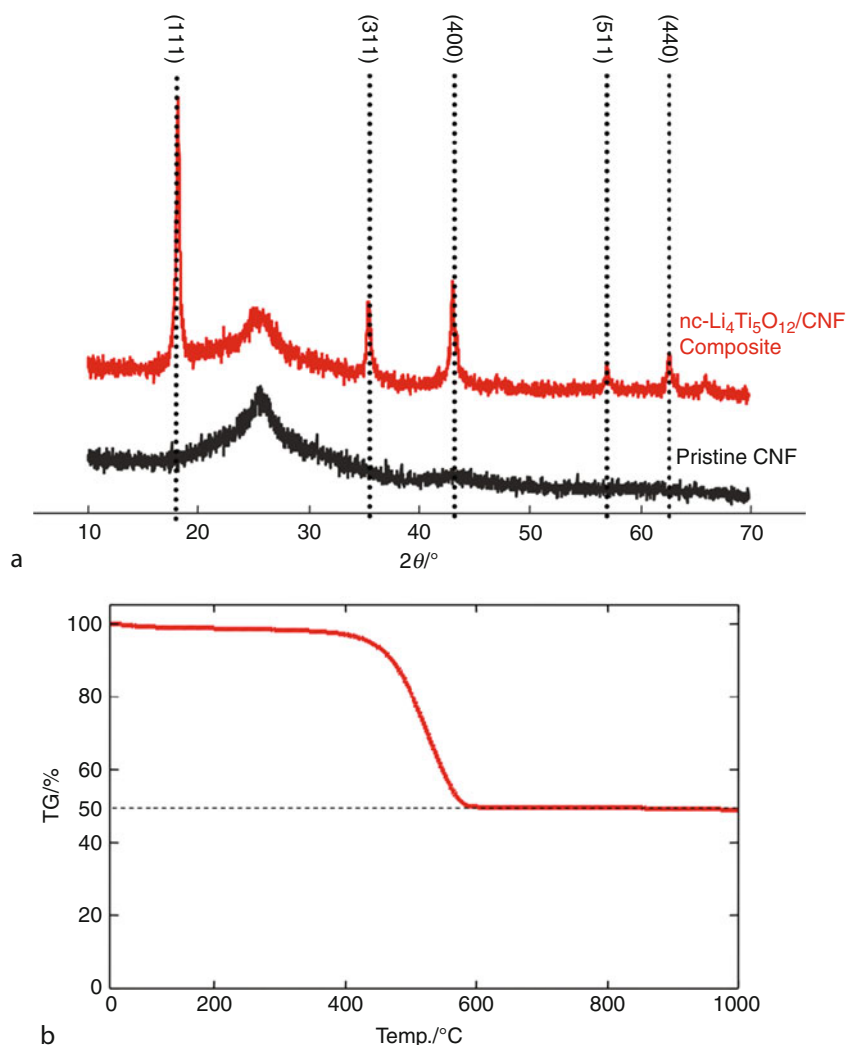
Electrochemical Supercapacitors and Hybrid Systems. Figure 14

Nano-structure model and high power performance of super-high-rate nano-crystalline Li₄Ti₅O₁₂ nested and grafted onto carbon nanofiber. **(a)** Schematic illustration for the two-step formation procedure of the nc-Li₄Ti₅O₁₂/CNF composite. **(b)** Maximum C rate values of the nc-Li₄Ti₅O₁₂/CNF composite and various Li₄Ti₅O₁₂ materials in literatures reported so far

particles in the composite are electrochemically active, meaning that ionic and electric paths are fully established in the composite. The rate capability of the obtained composite is shown in Fig. 18b.

Even at a high rate of 300 C, the composite shows reversible capacity of 158 mAh g⁻¹ per Li₄Ti₅O₁₂ which

correspond to 95% of the capacity obtained at 1 C. Such an excellent rate-capability indicates that the optimized nano-structure of the nc-Li₄Ti₅O₁₂/CNF composites as observed in HR-TEM image well overcome the inherent problems of Li₄Ti₅O₁₂ materials like poor Li⁺ diffusivity and poor electronic conductivity. Probably,



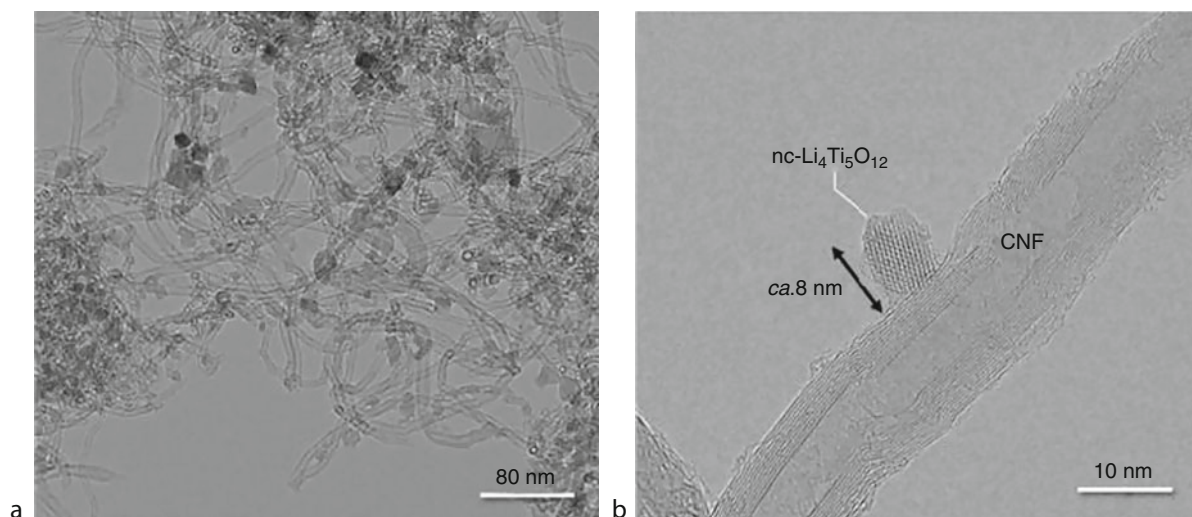
Electrochemical Supercapacitors and Hybrid Systems. Figure 15

Crystallinity and content of $\text{nc-Li}_4\text{Ti}_5\text{O}_{12}$ in the composite. **(a)** XRD patterns of the $\text{nc-Li}_4\text{Ti}_5\text{O}_{12}/\text{CNF}$ composite and pristine CNF. **(b)** TG curve of the $\text{nc-Li}_4\text{Ti}_5\text{O}_{12}/\text{CNF}$ composite at 1°C min^{-1} under air. The residual weight ratio corresponds to the content of $\text{nc-Li}_4\text{Ti}_5\text{O}_{12}$ in the composite

this is because the nano-crystallized $\text{Li}_4\text{Ti}_5\text{O}_{12}$ and $\text{Li}_4\text{Ti}_5\text{O}_{12}/\text{CNF}$ junctions lead to facile ionic diffusion and electronic conduction, respectively. The cycleability of the composite is shown in Fig. 18c. Even after 9,000 cycles, 90% of the initial capacity is maintained, showing that the composite is electrochemically stable. The result strongly suggests that the

aggregation and detachment of the $\text{nc-Li}_4\text{Ti}_5\text{O}_{12}$ particles hardly happen when they are operated at high-rate charge–discharge for a prolonged cycling.

Figure 19 shows Ragone plots obtained from charge–discharge measurements as laminate-type cell of the hybrid capacitor system ($(\text{nc-Li}_4\text{Ti}_5\text{O}_{12}/\text{CNF})/\text{LiBF}_4\text{-PC/AC}$). The charge–discharge was performed



Electrochemical Supercapacitors and Hybrid Systems. Figure 16

Nano-structure of the $\text{nc-Li}_4\text{Ti}_5\text{O}_{12}$ and CNF. (a) A bird's eye view of a HR-TEM image of the $\text{nc-Li}_4\text{Ti}_5\text{O}_{12}$ /CNF composite and (b) a worm's eye view focusing on the junction of the $\text{nc-Li}_4\text{Ti}_5\text{O}_{12}$ particle on CNF graphene surface



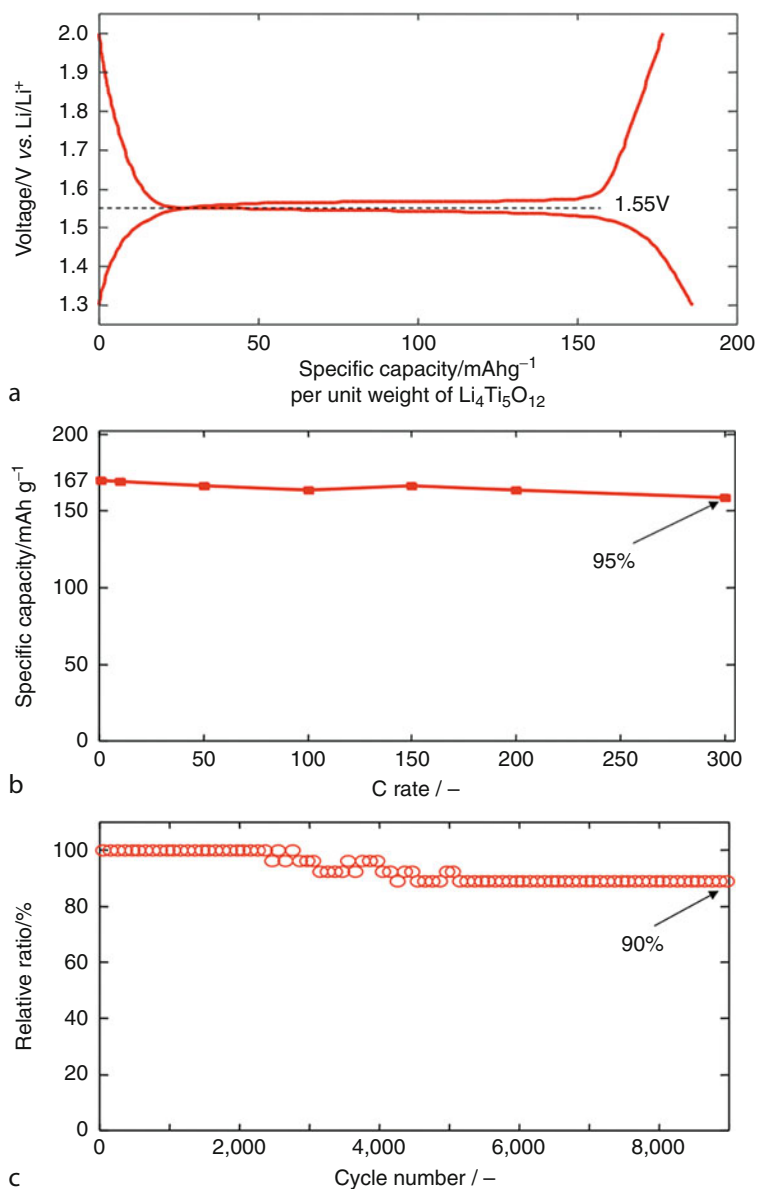
Electrochemical Supercapacitors and Hybrid Systems. Figure 17

Closer view of a HR-TEM image of multiple attachments of $\text{nc-Li}_4\text{Ti}_5\text{O}_{12}$ inside and outside wall of the CNF composite

between 1.5 and 3.0 V at various current densities ranging $0.2\text{--}30\text{ mA cm}^{-2}$ ($0.18\text{--}26.8\text{ A g}^{-1}$). For comparison, a conventional EDLC system (AC/TEABF₄-PC/AC) was also assembled and measured between 0 and 2.5 V. In a low power density range of $0.1\text{--}1\text{ kW L}^{-1}$, the hybrid capacitor shows the energy density as high as $28\text{--}30\text{ Wh L}^{-1}$, which is a value comparable to that of the Li-ion capacitors [31]. Even at a high power of 6 kW L^{-1} , the energy density of the hybrid capacitor remains at 15 Wh L^{-1} which is double that of the conventional EDLC system (AC/AC). The result reveals that the capacitor system can provide higher energy as compared with conventional EDLCs not only in the low power density region ($0.1\text{--}1\text{ kW L}^{-1}$) but also in the high power density region ($1\text{--}6\text{ kW L}^{-1}$). Accordingly, this configuration of capacitor system is anticipated as an energy device utilizable for both high energy and high power applications.

Future Perspectives

Hybrid capacitor systems are the promising approach to meet the goal to effectively increase the energy

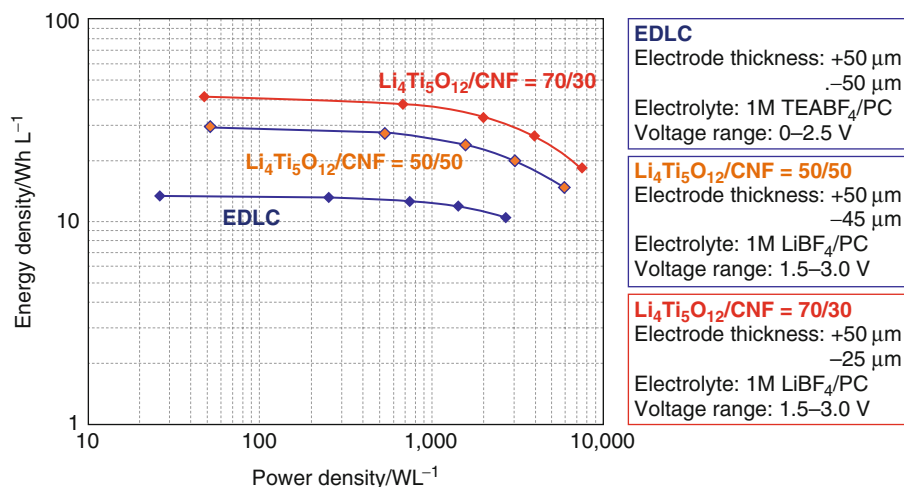


Electrochemical Supercapacitors and Hybrid Systems. Figure 18

Electrochemical properties of the nc-Li₄Ti₅O₁₂/CNF composite. (a) Charge-discharge curve of nc-Li₄Ti₅O₁₂/CNF composite between 1.0 and 3.0 V versus Li/Li⁺ at 1 C. (b) Rate capability of the nc-Li₄Ti₅O₁₂/CNF composite in a C-rate ranges from 1 C to 300 C. (c) Cycleability of nc-Li₄Ti₅O₁₂/CNF composite between 1.0 and 3.0 V versus Li/Li⁺ at 20 C. 1 M LiBF₄/EC + DEC(1:1) was used as the electrolyte

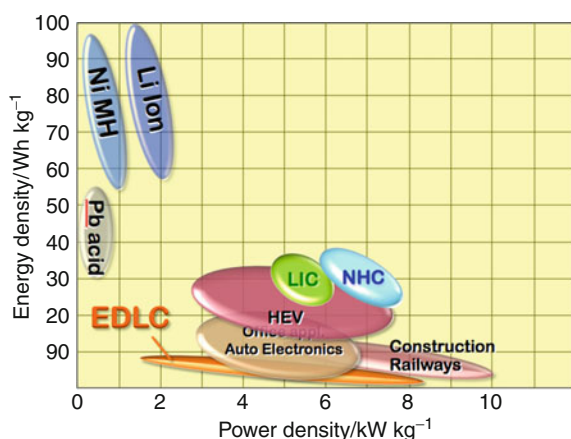
density. The investigation to develop hybrid capacitors has been initiated by “Li-ion capacitors (LIC).” And, now “Nanohybrid capacitor (NHC)” certainly achieves as high energy density as Li-ion capacitors with higher

stability, higher safety, and higher productivity (see Fig. 20). Both LIC and NHC attain three-times higher energy density and are regarded as the next-generation supercapacitor systems.



Electrochemical Supercapacitors and Hybrid Systems. Figure 19

Ragone plots of Nanohybrid capacitor systems ((nc-Li₄Ti₅O₁₂/CNF)/AC) with two different Li₄Ti₅O₁₂ loadings (50% and 70%) and conventional EDLC system (AC/AC) all in 1 M LiBF₄/PC



Electrochemical Supercapacitors and Hybrid Systems. Figure 20

Figure 20

Performances of lithium-ion capacitor (LIC) and nanohybrid capacitor (NHC) in Ragone plot

Bibliography

- Amatucci GG, Badway F, Du. Pasquier A, Zheng T (1999) In: Abstract of 196th meeting of the Electrochemical Society, Honolulu, p 122
- Amatucci GG, Badway F, Pasquier ADu, Zheng T (2001) An asymmetric hybrid nonaqueous energy storage cell. *J Electrochem Soc* 148:A930
- Azaïs P, Tetrais F, Caumont O, Depond JD, Lejosne J (2009) Ageing study of advanced carbon/carbon ultracapacitor cells working in various organic electrolytes. In: Abstract of ISEE'Cap 09, p 19
- Bai Y, Wang F, Wu F, Wu C, Bao L (2008) Influence of composite LiCl-KCl molten salt on microstructure and electrochemical performance of spinel Li₄Ti₅O₁₂. *Electrochim Acta* 54:322
- Baldsing WG, Puffy NW, Newnham RH, Pandolfo AG (2007) High-energy asymmetric nickel-carbon supercapacitors. In: Proceedings advanced automotive battery and ultracapacitor conference, Long Beach
- Burke A (2000) Ultracapacitors: why, how, and where is the technology. *J Power Sources* 91:37
- Burke A (2007) R&D considerations for the performance and application of electrochemical capacitors. *Electrochim Acta* 53:1083
- Chen CH, Vaughey JT, Jansen AN, Dees DW, Kahaian AJ, Goacher T, Thackeray MM (2001) Studies of Mg-substituted Li_{4-x}Mg_xTi₅O₁₂ spinel electrodes (0x1) for lithium batteries. *J Electrochem Soc* 148:A102
- Hao Y, Lai Q, Xu Z, Liu X, Ji X (2005) Synthesis by TEA sol-gel method and electrochemical properties of Li₄Ti₅O₁₂ anode material for lithium-ion battery. *Solid State Ionics* 176:1201
- Hatozaki O (2008) Lithium ion capacitor: electrode materials and cell performance. In: Proceedings advanced capacitor world summit 2008, San Diego (2008)
- Huang J, Jiang Z (2008) The preparation and characterization of Li₄Ti₅O₁₂/carbon nano-tubes for lithium ion battery. *Electrochim Acta* 53:7756
- Huang S, Wen Z, Zhu X, Gu Z (2004) Preparation and electrochemical performance of Ag doped Li₄Ti₅O₁₂. *Electrochem Commun* 6:1093
- Jansen AN, Kahaian AJ, Kepler KD, Nelson PA, Amine K, Dees DW, Vissers DR (1999) High-rate nano-crystalline Li₄Ti₅O₁₂

- attached on carbon nano-fibers for hybrid supercapacitors. *J Power Sources* 81–82:902
14. Kazaryan SA, Kharsov GG, Litvinenko SV, Kogan VI (2007) Self-discharge related to iron ions and its effect on the parameters of HES $\text{PbO}_2/\text{H}_2\text{SO}_4/\text{C}$ Systems. *J Electrochem Soc* 154:A751
 15. Kim J, Cho J (2007) Spinel $\text{Li}_4\text{Ti}_5\text{O}_{12}$ nanowires for high-rate Li-Ion intercalation electrode. *Electrochem Solid State Lett* 10:A81
 16. Koetz R, Carlen M (2000) Principles and applications of electrochemical capacitors. *Electrochim Acta* 45:2483
 17. Laforge A, Simon P, Fauvarque JF, Mastrangostino M, Soavi F, Sarrau JF, Lailler P, Conte M, Rossi E, Saguatti S (2003) Activated carbon/conducting polymer hybrid supercapacitors. *J Electrochem Soc* 150:A645
 18. Machida K, Suematsu S, Ishimoto S, Tamamitsu K (2008) High-voltage asymmetric electrochemical capacitor based on polyfluorene nanocomposite and activated carbon. *J Electrochem Soc* 155:A970
 19. Naoi K, Ishimoto S, Ogihara N, Nakagawa Y, Hatta S (2009) Encapsulation of nanodot ruthenium oxide into KB for electrochemical capacitors. *J Electrochem Soc* 156:A52
 20. Naoi K, Simon P (2008) New materials and new configurations for advanced electrochemical capacitors. *Interface* 17:34
 21. Nikkei Electronics (2008) LIC prevails EDLCs?! 991:77
 22. Ohzuku T, Ueda A, Yamamoto N (1995) Zero-strain insertion material of $\text{Li}[\text{Li}_{1/3}\text{Ti}_{5/3}]\text{O}_4$ for rechargeable lithium cells. *J Electrochem Soc* 142:1431
 23. Pandolfo AG, Hollenkamp AF (2006) Carbon properties and their role in supercapacitors. *J Power Sources* 157:11
 24. Plitz I, Dupasquier A, Badway F, Gural J, Pereira N, Gmitter A, Amatucci GG (2006) Nanohybrid capacitor: The next generation electrochemical capacitors. *Appl Phys A* 82:615
 25. Scharner S, Weppner W, Schmind-Beurmann P (1999) Evidence of two-phase formation upon lithium insertion into the $\text{Li}_{1.33}\text{Ti}_{1.67}\text{O}_4$ spinel. *J Electrochem Soc* 146:857
 26. Shen CM, Zhang XG, Zhou YK, Li HL (2002) Preparation and characterization of nanocrystalline $\text{Li}_4\text{Ti}_5\text{O}_{12}$ by sol-gel method. *Mater Chem Phys* 78:437
 27. Shu J (2008) Study of the interface between $\text{Li}_4\text{Ti}_5\text{O}_{12}$ electrodes and standard electrolyte solutions in 0.0–5.0 V. *Electrochem Solid State Lett* 11:A238
 28. Simon P, Gogotsi Y (2008) Materials for electrochemical capacitors. *Nat Mater* 7:845
 29. Takai S, Kamata M, Fujiine S, Yoneda K, Kanda K, Esaka T (1999) Diffusion coefficient measurement of lithium ion in sintered $\text{Li}_{1.33}\text{Ti}_{1.67}\text{O}_4$ by means of neutron radiography. *Solid State Ionics* 123:165
 30. Thackeray MM (1995) Structural considerations of layered and spinel lithiated oxides for lithium ion batteries. *J Electrochem Soc* 142:2558
 31. Yoshino A, Tsubata T, Shimoyamada M, Satake H, Okano Y, Mori S, Yata S (2004) Development of a lithium-type advanced energy storage device. *J Electrochem Soc* 151:A2180
 32. Yu H, Zhang X, Jalbout AF, Yan X, Pan X, Xie H, Wang R (2008) High-rate characteristics of novel anode $\text{Li}_4\text{Ti}_5\text{O}_{12}$ /polyacene materials for Li-ion secondary batteries. *Electrochim Acta* 53:4200
 33. Zhou JH, Sui ZJ, Li P, Chen D, Dai YC, Yuan WK (2006) Structural characterization of carbon nanofibers formed from different carbon-containing gases. *Carbon* 44:3255

Electronic Waste and Its Regulation

SHIRLEY THOMPSON

Natural Resources Institute, University of Manitoba, Winnipeg, MB, Canada

Article Outline

Glossary
 Definition of the Subject and Its Importance
 Introduction
 Regulation of Electronic Waste
 Future Directions
 Bibliography

Glossary

Design for environment (DfE) The various design approaches geared at reducing the overall environmental impact of a product, process, or service, considering environmental impacts across the product's life cycle.

Ecycling The process of recycling the components in used or discarded electronic equipment.

Electronic waste, e-waste and waste electrical and electronic equipment (WEEE) Old, end-of-life, or discarded appliances using electricity and includes computers, consumer electronics, fridges, etc. which have been disposed of by their original users.

Extended producer responsibility (EPR) Entails making manufacturers responsible for the entire life cycle of the products and packaging they produce based on the "polluter pays principle."

Persistent and bioaccumulative toxins (PBT) Compounds that are resistant to environmental degradation. Because of this, they have been observed to persist in the environment, bioaccumulate in human and animal tissue, and to have potential significant impacts on human health and the environment.

Toxicity characteristic leaching procedure (TCLP)

A soil sample extraction method for chemical analysis employed as an analytical method to simulate leaching through a landfill.

Definition of the Subject and Its Importance

Electronic waste or “E-waste” is broadly defined as all surplus electronics. Electronic wastes are computers, television sets, mobile phones, etc., destined for reuse, resale, salvage, recycling, or disposal. Computers are a complex waste problem due to their volume, toxicity, and disposable, rather than recyclable, product design. Computers quickly become obsolete as technology rapidly changes. Everyday more than 160,000 obsolete computers and televisions are generated in the USA, weighing almost 3.2 thousand metric tons (3.5 thousand US tons). Electronic waste is about 4% of the total solid waste stream in North America but is growing two to three times faster than any other waste stream (i.e., paper, yard waste). For every new computer coming on the market there is an obsolete computer [1]. In 1998, of the 20 million computers taken out of service in the USA, only 2.3 million (slightly more than 10%) were recycled or E-cycled, mostly from large businesses and institutions. As well, 75% of “obsolete” computers are being stockpiled in people’s homes [2].

Introduction

With rapid improvements in technology, most computers are disposed of within 2 years, rather than at the end of their functional life cycle of approximately 10 years [3]. Each new innovation in computer technology doubles its obsolescence rate, without increasing its repairability or recyclability. The consumer, targeted by advertising about the latest computers, finds buying a new computer cheaper and more convenient than upgrading their old one. Computers that cannot be upgraded become waste.

Electronic waste is a complex waste, made up of more than 30 compounds, many of which are hazardous substances. The presence of hazardous substances in computers makes landfilling or incinerating environmentally unsuitable for discarded computers. The use of toxic metals and other chemicals in computers results in environmental and health risks when computers are manufactured, incinerated, landfilled,

burned, or melted down during recycling [4, 5]. E-waste emits billions of kilograms of toxic materials into the air, water, and soil, most of which are persistent and bioaccumulative toxins (PBTs). A cathode ray tube (CRT), the most common type of computer monitor until recently, typically contains four pounds of lead to protect users from the tubes’ X-rays [2]. Lead is a teratogen and reproductive hazard. As well, lead causes damage to the central and peripheral nervous systems, blood system, and kidneys in humans. Circuit boards and batteries also contain lead, in addition to smaller amounts of mercury and hexavalent chromium, nickel, bromine, and zinc. In 2000, 90% of e-waste had no removal of components containing toxic mercury switches or nickel batteries before landfilling or incineration in Europe and North America. With these harmful chemicals, many consumer electronics fail the Toxicity Characteristic Leaching Procedure (TCLP) [6]. Leaching of heavy metals and brominated compounds occurs during landfilling, contaminating the soil and groundwater beneath a site and the surrounding area. For example, the European Commission estimates that consumer electronics constitute 40% of the lead found in landfills. Moreover, the workers in e-waste recycling operations face dangerous health and environmental risks. Hence the concerns about the trade, the technology in practice, and the existing poor disposal practices of e-waste in North America and many nations.

Increased regulation of electronic waste and concern over the environmental harm, which can result from toxic electronic waste has raised disposal costs. Regulation creates an economic disincentive to dispose of residue, unless strict restrictions enforce proper treatment. In extreme cases, brokers and others calling themselves recyclers export unscreened electronic waste to developing countries, avoiding the expense of removing items like bad cathode ray tubes (the processing of which is expensive and difficult).

Hazardous e-waste from developed countries (e.g., Western European countries, USA, Canada, and Japan) has a history of being recycled in developing countries and in Eastern Europe. These products are exported when the state has banned various metals in their landfills and incinerators, rather than pay the higher labor and environmental costs for recycling locally [7]. As much as 80% of all recycled e-waste in the USA were

shipped to Asia (e.g., China, Cambodia, India, Indonesia, Malaysia, Manila, and Pakistan) all of which have poor regulation and management of e-waste [8, 9, 10]. This shipping to places with low labor costs and environmental regulation occurs for the EU as well but the new WEEE-IT (Information Technology) regulation at least requires that their e-waste be recycled at facilities having as high or higher environmental and safety standards as in the EU.

A few demanufacturing operations exist in North America but are typically manned by unskilled or semiskilled labor such as prisoners, former welfare recipients, or mentally disabled individuals. These workers are exposed to hazardous conditions and often do not earn a living wage [11]. For example, 250 inmates, at the Atwater federal US penitentiary's recycling plant process 450 computers daily. The inmates receive only US \$0.23–\$1.15 per hour for this work and work without safety equipment, such as gloves, dust masks to reduce exposures to toxic lead, and barriers to prevent physical risks from cuts and lead exposure. Hammers are often used to break cathode ray tubes (CRT) rather than a sealed crushing machine, which is typical in commercial recycling, carrying out this work [12].

Regulation of Electronic Waste

Extended producer responsibility (EPR) regulations in Europe hold producers who are legally responsible in four ways for their products. They are responsible for having: (1) Economic responsibility to pay all or a portion of end-of-life management, (2) Physical responsibility to take possession of their products after the consumer discards it, (3) Information responsibility to provide mandatory product labeling such as component or material lists, and (4) Financial liability for environmental damage and clean-up costs from disposal of hazardous products [6, 13]. Generally, EPR policies have three characteristics: (1) a focus on end-of-life waste management to encourage redesign for environment, (2) a shift of physical and/or financial responsibilities from taxpayer or consumer to producer, and (3) an explicit target for waste reduction. Mandated programs “force producers to get involved in managing material streams” [14]. Consumers of WEEE products must have the opportunity to return

waste items, without charge, to collection facilities with the producer responsible for product recovery.

Extended producer responsibility extends the “polluter pays principle” to consumer products, requiring “that the polluter should, in principle, bear the cost of pollution, with due regard to the public interest and without distorting international trade and investments” [15]. EPR policy places responsibility for a product's end-of-life environmental impacts on the original producer and seller of that product requiring producers: to take back their products and to meet recycling rate targets. This approach provides incentives for producers to optimize their internal logistics and recycling systems to minimize costs, create value, and integrate eco-design into business routines. Those changes should include improving product recyclability and reusability, reducing material usage and downsizing products, lower energy consumption and greenhouse gas production, reduce dependency on virgin materials, spur new business enterprises, generate new job opportunities, and provide financial and savings to companies improving their design, production, and distribution processes [16]. However, the ability to gain competitive advantage is restrained by the domination of collective compliance structures across the EU, over whose costs producers regard as difficult to manage or influence.

Despite new EU shipment rules banning the export of hazardous materials in waste electrical equipment to non-OECD countries from January 2002, a few exporters have transferred over two million old computer monitors and one million old TVs out of the EU for reuse. For example, of an estimated 121,000 metric tons (133,000 US tons) of computer and telecommunications equipment exported from the UK for re-use or recycling in 2003, 21,000 metric tons (23,000 US tons) was “grey market” equipment being exported without the full necessary documentation. This equipment is rarely tested or repaired before export to countries where dismantling takes place in unsafe conditions, or the equipment is simply dumped. Without criteria for the reuse of IT appliances (e.g., testing, minimal capacity, etc.) this environmental injustice can continue. Very little consumer-based electronic waste has true reuse value, and claims of export for reuse should be scrutinized carefully. Although Canada prohibits waste export to non-OECD nations, this occurs disguised as computer donations for reuse.

According to WEEE IT EU regulations, an average of 75% recovery by weight and of 65% component, material, and substance reuse and recycling must be achieved by December 31, 2006 [17]. These targets appear feasible, even low, based on pre-WEEE-IT programs surpassing these targets. Prior to WEEE, several member states within the EU (Belgium, the Netherlands, Sweden, Norway, and Switzerland) adopted national regulations and management schemes for e-wastes [6], with programs operating at a recycling rate (including energy recovery) of approximately 80–97% with 97% achieved by Switzerland [18]. Although high targets are being set and a good recycling/reuse rate is achieved, still a significant amount of waste (10–20%) is not recovered for recycling and reuse, presumably ending up in landfills or incinerators, releasing toxic chemicals. That some 91,000 metric tons (100,000 US tons) are incinerated or landfilled each year means computer waste has not been prevented and mass consumption of this disposal product continues to occur [19].

The USA and Canada have, rarely, if ever, mandated strict guidelines for product manufacturing, which includes both how the product is produced and what types of materials are used – perhaps as a result of the lack of political pressures to develop new waste management strategies amidst relatively plentiful landfill accommodations. Policy makers in North America are reluctant to require sustainable product management. A few companies, such as HP and Dell, do voluntarily take back their products, but at high costs that many consumers are unwilling to pay or have severe restrictions that limit acceptance to consumption of their new materials. The recycling rates of voluntary computer take-back programs in the USA and Canada conducted by a few companies, such as Dell, HP, Sony amount to only 14%, according to the most recent available figures, due to cost and acceptance of e-waste in landfills [2]. Many economically successful reuse and recycling programs are in effect [15, 20, 21, 22], even with producers' focus on productivity and the bottom-line. Hewlett-Packard (HP) was the first computer corporation to have a “tear it down” operation to accompany its successful “build it up” operation called Product Recycling Solutions (PRS), launching a take-back program in 1992 [15] in which HP committed to reuse or donate or environmentally sound recycle their

products [7]. HP limits their financing to the pickup and transport of their printers and cartridges, requiring customers to pay \$25–\$50 Canadian for each e-waste component (CPU, monitor, scanner, server). HP witnessed numerous economic and market-share benefits from reducing resource and energy consumption of their products and production processes. An 87% landfill diversion rate reduced landfill costs, and reusing items accounted for \$2.9 million Canadian in benefits for HP in addition to \$2 million revenue for selling recycled materials in 2006 [23]. By applying reuse at HP, 15% of defects were reduced and productivity increased by 57% [22]. However, it should be noted that although the company avoided 16,000 metric tons (18,000 US tons) of carbon dioxide (CO₂) in 2006 through energy efficiency measures their emissions in 2006 were approximately 100x that at roughly 1,600,000 metric tons (1,800,000 US tons) of carbon dioxide [23]. The Electronic Product Stewardship Canada (2006) reports 91,000 metric tons (100,000 US tons) of electronics reach end of life in Canada annually and only 11% of this is currently recycled.

Future Directions

Policies can encourage redesign for reuse and upgrading for longevity, as well as recycling: Computers can be redesigned to have less material (de-materialization), facilitate upgrading through modularity, last longer, expend less energy and pollution in their production and use and that disassemble for reuse and recycling with ease after their useful life. Design for the Environment (DfE) promotes consideration of design issues related to environmental and human health over the life cycle of the product. Material selection, energy use, extended component life cycles, disassembly, reprocessing, remanufacturing must all be considered [24]. DfE has eight axioms: (1) manufacture without producing hazardous waste, (2) use clean technologies, (3) reduce product chemical emissions, (4) reduce product energy consumption, (5) use non-hazardous recyclable materials, (6) use recycled material and reused components, (7) design for ease of disassembly, and (8) product reuse or recycling at end of life [25]. Principles of ecodesign require products to be flexible, reliable, durable, adaptable, modular, dematerialized, and reusable. As applies to computers, improving environmental factors can

improve performance as well: work on optimization of environmental performance often spurs creativity and brings innovative design solutions into the product development process which allows improvement of its overall performance [26].

Policies can encourage reuse, as well as, recycling. Recycling is the series of activities by which discarded materials are collected, sorted, processed, and used in the production [27]. Although recycling reduces virgin material use, it still requires additional energy to be used to reform them into manufactured products [27]. Silicon Valley Toxics Coalition (2001) estimates that the cost of recycling computers ranges from \$10 to \$30 per unit. While this is less expensive than the estimated \$25–50 per unit cost for disposal, the value of recycled materials in computers is \$2.00–\$3.00.

In contrast, remanufacturing, reconditioning, and repair should save money for the producer and reduce costs for the consumer. Over each computer's life cycle, reselling or upgrading computers saves 5–20 times more energy than recycling and would provide a public good if provided to organizations for educational purposes. For example, public schools have large needs for computers at 28.7 million computers in USA, which may be met by reusing or remanufactured computers, as recent state and local budget cuts otherwise severely limit the affordability of having computers in the classroom [28]. Reuse can include repair, recondition, or remanufacture. Also, it is environmentally preferable to reuse computers an extra 2 or 3 years than to recycle their components. Reducing the amount of computer waste relies heavily upon the reuse of systems that may be out of date, but fully functional [3]. Remanufacturing is considered the best strategy as the embodied energy of virgin production is maintained, preserving the intrinsic "added value" of the product for the manufacturer and enables the resultant products to be sold "as new" with updated features if necessary [27]. Components and materials from the end-of-life durable products can often be refurbished to substitute for virgin parts to be used as spares or in remanufacturing [29]. Upgrading individual components of a computer at the design stage, such as the hard drive, software, or processor, costs less than a new computer [3]. Lund, R. (1985) estimates that a remanufacturing product only requires 20–25% of the energy used in its initial formation and can be

resold into lower-priced markets, typically costing 60% of the original production cost [27, 30]. However, the lack of a credible and stable demand for recycled or refurbished products is a barrier to remanufacturing [27, 31]. The noncommercial computer refurbishing field is a minuscule part of the overall computer recycling industry in the USA and Canada.

Designing for the environment (DfE) and implementing green production processes makes economic sense when life cycle environmental costs are required to be paid by the producer. DfE has producers considering "at the development phase of a products life cycle, the environmental impacts through enhancing the product design which includes resource consumption, both in material and energy terms and pollution prevention." Environmental considerations in product design include: waste minimization, reuse or recyclability, material conservation, pollution reduction, lower toxicity and "eco-design" [15, 32]. Six principles of Design for the Environment (DfE) are efficiency, appropriateness, sufficiency, equity, systems, and scale [33]. However, without extended producer responsibility internalizing environmental costs, manufacturers are typically unwilling to improve product and environmental quality if it costs more or compromises costs, quality, and scheduling constraints [24]. Firms are only willing to spend 0.1% more on a product to improve environmental quality [34]. Thus, usually design takes on the character of being a problem-solving activity for a component rather than product development from its conceptual stage [35, 36].

Bibliography

Primary Literature

1. Thompson S, Oh S (2006) Do sustainable computers result from design for environment and extended producer responsibility?: analyzing e-waste programs in Europe and Canada. ISWA/NVRD world congress 2006 conference proceedings. ISWA/NRVD, Copenhagen, Denmark, pp 145–161
2. Gattuso DJ (2005) Mandated recycling of computers: a lose-lose proposition. Competitive Enterprise Institute (CEI)
3. Williams E, Sasaki Y (2003) Energy analysis of end-of-life options for computers: resell, upgrade, recycle. In: Proceedings of the 2003 IEEE international symposium on electronics and the environment, San Francisco, CA, pp187–192
4. Lincoln JD et al. (2005) Environmentally benign materials for electronics: A review current development and merging

- technologies. In Proceedings. International symposium on advanced packaging materials: processes, properties and interfaces. IEEE, Piscataway, NJ
5. Schmidt CW (2002) E-Junk explosion. *Environ Health Perspect* 110(4):232–235
 6. Sachs N (2006) Planning the funeral at the birth: extended producer responsibility in the European Union and the United States. *Harv Environ Law Rev* 30:51–98
 7. McCarthy JE (2002) Recycling computer and electronic equipment: legislative and regulatory approaches for “E-Waste,” Congressional Research Service, pp 1–21
 8. Iles A (2004) Mapping environment justice in technology flows: computer waste impacts in Asia. *Glob Environ Polit* 4(4):76–105
 9. Terazono A et al (2006) Current status and research on E-waste issues in Asia. *J Mater Cycle Waste Manage* 8:1–12
 10. Yanez L et al (2002) Overview of human health and chemical mixtures: problems facing developing countries. *Environ Health Perspect* 10(110):901–909
 11. Price JL (1999) Reclaiming end-of-life cathode ray tubes (CRTs), and electronics: a Florida update, hazardous material Management conference, Tucson, Arizona. 15 Nov 1999
 12. Perry TS (2005) Recycling behind bars. *IEEE Spectr* 42(6):10–13
 13. Toffel MW (2002) The regulatory and judicial roots of product stewardship in the United States. Haas School of Business/University of California, Berkeley
 14. McKerlie K, Knight K, Thorpe B (2006) Advancing extended producer responsibility in Canada. *J Cleaner Prod* 14:616–628
 15. Schwartz BJ, Gattuso DJ (2002) Extended producer responsibility: reexamining its role in environmental progress. Reason Public Policy Institute, Los Angeles, pp 1–65
 16. Quinn L (2006) Extended producer responsibility: new direction for waste management. *Eco-Journal* (March/April)
 17. Nakajima N, Vanderburg WH (2005) A failing grade for WEEE take-back programs for information technology equipment. *Bull Sci Technol Soc* 25(6):507–517
 18. Vossenaar R, Santucci L, Ramungul N (2006) Environmental requirements and market access for development countries: the case of electrical and electronic equipment. Trade and environment review. United Nations, New York
 19. Lin CK, Yan L, Davis AN (2002) Globalization, extended producer responsibility and the problem of discarded computers in China: an exploratory proposal for environmental protection. *Georgetown Int Environ Law Rev* 14:525
 20. Bosch J (2002) Maturity and evolution in software product lines: approaches, artifacts and organization. In: Proceeding of the second software product line conference
 21. Fiksel J (2001) Emergence of a sustainable business community. *Pure Appl Chem* 73:1265–1268
 22. Tomer A, Kimchi E (2004) Evaluating software reuse alternatives: a model and its application to an industrial case study. *IEEE Trans Software Eng* 30(9):601–612
 23. Hewlett Packard (2007) HP global citizenship report: waste and recycling. <http://www.hp.com/hpinfo/globalcitizenship/gcreport/operations/waste.html>. Accessed 27 Mar 2007
 24. Armstrong T (1997) Design for sustainability, sustainable consumption and production. IndEco Strategic Consulting, Ottawa
 25. Hill B (1993) Industry's integration of environmental product design. In: 1993 IEEE international symposium on electronics and the environment. IEEE, Virginia
 26. Thompson S, Sawyer J, Bonam R, Valdivia JE (2009) Building a better waste generation model: validating models with recovery rates at 35 Canadian landfills. *Waste Manage* 29:2085–2091
 27. King AM, Burgess SC, Ijomah W, McMahon CA (2005) Reducing waste: repair, recondition, remanufacture or recycle?. *Sust Dev*
 28. Lynch J (2004) Baseline study of noncommercial computer reuse in the United States. In: International symposium on electronics and the environment (IEEE). CompuMentor, San Francisco, pp 189–194
 29. Toffel MW (2004) Strategic management of product recovery. *Calif Manage Rev* 46(2):120–141
 30. Lund R (1985) Remanufacturing: the experience of the United States and implications for developing countries, UNDP project management report 2, World Bank Technical Paper No 31, pp 24–34. In: King AM, Burgess SC, Ijomah W, McMahon CA (2005) Reducing waste: repair, recondition, remanufacture or recycle? J Sustain Develop
 31. Bras B, McIntosh MW (1999) Product, process and organizational design for remanufacture – an overview of research. *Robot Comput-Integr Manuf* 15:167–178
 32. Calcott P, Walls M (2006) Can downstream waste disposal policies encourage upstream “design for environment”? The American Economic Review: Papers and Proceeding of the One Hundred Twelfth Annual Meeting of the American Economic Association 90 (2), 233–237
 33. Flecher K, Dewberry E (2002) Demi: a case study in design for sustainability. *Int J Sustain High Educ* 3(1):38–47
 34. Matthews HS, McMichael FC, Hendrickson CT, Hart DJ (1997) Disposition and end-of-life options for personal computers. Green design initiative technical report #97–10
 35. Fallman D (2003) Design-oriented human-computer interaction. *Designing Des* 5(1):225–232
 36. Nichol S, Thompson S (2007) Policy options to reduce consumer waste to zero: comparing product stewardship and extended producer responsibility for refrigerator waste. *Waste Manage Res* 25(3):227–233

Books and Reviews

- Boehm B (2003) Value-based software engineering. *ACM SIGSOFT Softw EngNotes* 28(2):1–11
- Bradner S, Mankin A (1995) The recommendation for the IP next generation protocol, Network Working Group
- Bras B (1997) Incorporating environmental issues in product design and realization. *Ind Environ* 20(1–2):1–19
- Business and Industry Advisory Committee (1998) Business and Industry Advisory Committee (BIAC) statement to the OECD EPR workshop 3 on extended and shared responsibility for products Washington, DC, 1–3 Dec 1998

- Cull K, Malins J (2003) Designing a sustainable future: a new approach to influence design practice. <http://www.ub.es/5ead/PDF/5%20/CullMalins.pdf>
- Farrell A (1996) Sustainability and the design of knowledge tools. *Technol Soc Mag* 15:17–21
- Hauschild M, Wenzel H, Alting L (1999) Life cycle design: enroute to the sustainable industrial culture? *Ann CIRP* 48(1):393–396
- Jacard M (2006) Mobilizing Producers toward Environmental Sustainability: The Prospects for Market-Oriented Regulations. In: Glen Toner (ed) *Sustainable production: building Canadian capacity*. UBC Press, Toronto
- Lilly D, Lofthouse V, Bhamra T (2005) Towards instinctive sustainable product use. In: *Sustainable creating the culture*, Aberdeen, 2–4th Nov
- Macauley M, Palmer K, Shih J (2003) Dealing with electronic waste: modeling the costs and environmental benefits of computer monitor disposal. *J Environ Manage* 68:13–22
- McDonough W, Braungart M (2002) Design for the triple top line: new tools for sustainable commerce. *Corp Environ Strategy* 9(3):251–258
- Merola L (2006) The COTS software obsolescence threat. In: Fifth international conference on commercial-off-the-shelf (COTS)-based software systems, pp 127–133
- Moffet J, Meyer S, Pezzack J (2006) Collaborative public policy for sustainable production: a broad agenda and a modest proposal. In: Toner G (ed) *Sustainable production: building Canadian capacity*. UBC Press, Toronto
- National Safety Council (1999) *Electronic product recovery and recycling baseline report: recycling of selected electronic products in the United States*. National Safety Council, Washington, DC, p 29
- Perdan S, Azapagic A, Clift R (2000) Teaching sustainable development to engineering students. *Int J Sustain High Educ* 1(3):267–279
- Sheehan B, Spiegelman H (2005) Extended producer responsibility policies in the United States and Canada. In: Scheer D, Rubik F (eds) *Governance of integrated product policy*. Greenleaf Publishing, Sheffield, pp 202–223
- Sheehan B, Spiegelman H (2005b) EPR in the US and Canada. *Resour Recycling* 2:18–21
- Sheehan B, Spiegelman H (2006) *Extended producer responsibility policies in the United States and Canada: history and status*. Greenleaf Publishing, Sheffield. Silicon Valley Toxics Coalition, Poison PCs and Toxic TVs
- Swanson M (2003) Canadian producer responsibility organizations. *E4 Partners*, pp 2–10. 15 Mar 2006
- Walls M (2006) *Extended producer responsibility and product design. Resource for the future*
- Wheeler D, Thomson K, Perkin M (2006) Sustainability, social capital and the Canadian ICT Sector. In: Glen Toner (ed) *Sustainable production: building Canadian capacity*. UBC Press, Toronto
- Xu CY, Morrison PD (2005) Sustainable product development and business performance: a product life cycle approach. ANZMAC conference: entrepreneurship, innovation and new product development

Energy and Water Interdependence, and Their Implications for Urban Areas

LIZ MINNE¹, ARKA PANDIT², JOHN C. CRITTENDEN³, MIROSLAV M. BEGOVIC⁴, INSU KIM⁴, HYUNJU JEONG², JEAN-ANN JAMES², ZHONGMING LU², MING XU⁵, STEVE FRENCH⁶, MUTHUKUMAR SUBRAHMANYAM⁶, DOUGLAS NOONAN⁷, MARILYN A. BROWN⁷, JESS CHANDLER⁷, YONGSHENG CHEN⁸, ERIC WILLIAMS⁹, REGINALD DESROCHES⁸, BERT BRAS¹⁰, KE LI¹¹, MICHAEL CHANG²

¹Environmental Engineering, Brook Byers Institute for Sustainable Systems, Georgia Institute of Technology, Atlanta, GA, USA

²Brook Byers Institute for Sustainable Systems and the Department of Civil & Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, USA

³Georgia Institute of Technology, Atlanta, GA, USA

⁴School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

⁵School of Natural Resources and Environment, University of Michigan, Ann Arbor, USA

⁶Center for Geographic Information Systems and the School of City and Regional Planning, Georgia Institute of Technology, Atlanta, GA, USA

⁷School of Public Policy, Georgia Institute of Technology, Atlanta, GA, USA

⁸Department of Civil & Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, USA

⁹Golisano Institute of Sustainability, Rochester Institute of Technology, Rochester, NY, USA

¹⁰George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

¹¹Department of Biological and Agricultural Engineering, University of Georgia, Atlanta, GA, USA

Article Outline

Glossary
Definition of Subject
Introduction
Methodologies
Current Infrastructure
Water and Energy Nexus

Beyond the Water and Energy Nexus
 Steps Toward a More Sustainable Future
 Future Directions
 Acknowledgments
 Bibliography

Glossary

Anthroposphere That part of the environment that is developed or adapted by humans.

Life cycle assessment (LCA) Analysis of the impact of a product or service through all the stages of its life including raw resource acquisition, material processing, fabrication, assembly, transportation, operation, disposal, recycling, repurposing, or reusing.

Low-impact development (LID) A way of managing stormwater runoff through decentralized water systems, such as green roofs and rainwater harvesting.

Material flow analysis (MFA) Analysis of how resources move through the industrial, consumer, and ecological sectors.

Resilience The ability for a structure to maintain its fitness and function in response to indigenous and exogenous stressors; it is characterized by four measures:

Rapidity. A measure of the capacity to contain losses or prevent further degradation in a timely manner.

Redundancy. A measure of the inherent substitutability.

Resourcefulness. A measure of the capacity to mobilize resources to restore functionality in the event of disruption.

Robustness. An ability of the system to withstand a given level of stress and/or demand.

Sustainability The development of an anthroposphere that exists within the means of nature. Sustainability requires that humans use no more resources than nature can renewably provide and generate only wastes that nature can assimilate in a timely manner and without overwhelming natural cycles.

Urban heat-island effect The phenomenon where the temperature in a developed area with a high concentration of buildings, roads, and other physical infrastructure is significantly higher than in the surrounding undeveloped areas.

Urban sustainability An assessment of the long-term viability of a place considering the physical infrastructure (such as water, sewer, transportation, and buildings), socioeconomic issues (such as human well-being, health, wealth, safety, and security), and natural environment (such as resources, wildlife, and ecosystem health).

Definition of the Subject

There are many definitions for sustainability. Mathis Wackernagel, creator of the ecological footprint concept, defined sustainability as “securing people’s quality of life within the means of nature” [1]. The United Nations’ World Commission on Environment and Development (the Brundtland Commission) defined sustainable development as “development that meets the needs of the present without compromising the ability of future generations to meet their own needs” [2]. Throughout this report, sustainability will be defined as the development of the anthroposphere within the means of nature. Here, the anthroposphere is the built environment or the environment that humans created for security, wealth generation, and comfort. For it to exist within the means of nature implies that the anthroposphere must use resources that nature provides, and generate only the kind and amount of waste that can be assimilated into the environment without overwhelming natural cycles. At the global scale, there are several examples that suggest that current development patterns are unsustainable. Presently worldwide, there are approximately 7 billion people using 14 Gt of materials [3]. With only 5% of the global material use being renewable, the extraction of natural resources is beyond what nature can reasonably supply. At the other end, an enormous amount of synthetic and potentially toxic materials are being introduced into the global material cycle.

To fully understand sustainability, one needs to take a whole system approach. In the traditional reductive engineering paradigm, each of the individual infrastructure components such as water, energy, and transportation are optimized separately. However, since the function of all the component elements depend on each other, a more optimal solution can occur if all the urban pieces are considered together. These interdependencies form an ecosystem of infrastructure that,

if functioning properly, provides a lasting basis for human enterprise. An example of the interdependencies and the need for them to be considered together is the water energy nexus. In all but the most primitive cultures, it takes water to create useful energy and energy to create useful water. This nexus requires a comprehensive understanding not just of power generation or of water resource management, but of both and how they connect and interact at temporal and spatial scales that are both large and small. And on top of these engineering relationships, environmental, social, and economical concerns must also be integrated.

Why is sustainability, particularly urban sustainability, so important? The United Nations Environmental Program (UNEP) chief, Klaus Toepfer, stated in 2005 that “cities pull in huge amounts of resources including water, food, timber, metals and people. They export large amounts of wastes including household and industrial wastes, wastewater and the gases linked with global warming. Thus their impacts stretch beyond their physical borders, affecting countries, regions and the planet as a whole. So the battle for sustainable development, for delivering a more environmentally stable, just, and healthier world, is going to be largely won and lost in our cities” [4]. The UNEP expects 64% of the world’s population to live in urban areas by 2040, a dramatic increase from 50.5% in 2010 [5]. For the developed countries like the United States, those in the European Union, and others, the share of urban population is projected to be 87% in the same time period [5]. With such an influx of people from rural areas to urban areas, cities must be prepared for this mass in-migration. With much of this movement expected in developing countries, those that adopt sustainability as a guiding framework for development could leapfrog other countries that are financially and systemically burdened with maintaining the status quo. For the developed world, a hybrid approach in which existing centralized systems are augmented by decentralized systems could help those cities remain competitive and viable. New or old, however, each city has different demographics, cultural values, fiscal and physical constraints, climate, and topology, and any solution to the demand for urban infrastructure – sustainable or not – will have to be uniquely tailored to consider these differences.

Introduction

Transmission grids are an integral piece of urban infrastructure. Understanding how electricity is used and finding ways to decrease the use of energy will be essential for developing more sustainable urban systems. Likely the greatest opportunities for improving efficiencies and increasing sustainability lies not in a single solution, but in how all the pieces – old technology and new – interconnect and work together. Transmission grids not only transmit energy to residences and workplaces, but also to water treatment systems, to industries, to street lights, to trains, and to the gasoline stations where cars are fueled, just to name a few. These dependencies are not unidirectional, however. Energy production and demand are also significantly dependent on these systems. Pipelines move natural gas, oil, and water to refineries and power plants, while trains and ships transport coal for use in heating, manufacturing, and power production. Adding to this is the emergence of affordable and technologically feasible renewable or decentralized energy options such as photovoltaics, wind power, and microturbines. The most productive, healthy, and sustainable cities of the future will be the ones that learn to use all these resources to build and operate wholly integrated energy systems that meet the short- and long-term economic, social, and environmental needs and aspirations of their residents and region.

Urban areas are complex interconnected systems analogous to ecological food webs. For an urban area to function properly, water, energy, and transportation systems need to be effectively designed to support the area’s desired land use. Similarly, the land-use pattern also should be governed according to the capacity of the area to host development. Beginning in the nineteenth century, these systems have increasingly been designed, built, and operated as isolated systems without considering how they interact. Water purveyors have been concerned about water distribution. Power companies examined energy production, distribution, and usage. Transportation planners examined ways to move people and goods safely and efficiently. Generally, they cared about optimizing their own processes to satisfy their own needs. However, in recent years, some cities have begun to realize that they could improve efficiency by considering how these different

systems interact and function together within the whole urban system, rather than as individual and independent parts. For example, energy generation requires water for steam production and cooling. The waste heat from energy production could be used to produce hot water, heat homes, or provide air-conditioning. In another example, a Pacific Northwest National Laboratory study in 2007 suggested that 73% of the United States' current light-duty vehicle fleet could be supported by the existing electric power infrastructure [6]. It is less clear, however, how the increased electricity demand required to electrify transportation would affect water demand. These are just two simple examples. Socioeconomics, policy, land use, and urban infrastructure form a complex system that makes up the city. When a holistic approach is adopted to analyze the urban infrastructure, many properties emerge that are not apparent when the components are analyzed individually. These emergent properties evolve from the economic and sociopolitical framework adopted by the stakeholders for the urban area. Hence, citizen capacity building, whereby informed citizens demand more sustainable infrastructure, is vital to driving the political agenda to improve the urban infrastructure.

A city is such a multifaceted and complex system that one cannot easily test the effect of certain decisions or actions. However, the impact of decisions can be estimated through modeling that simulates how agents change their behavior in reaction to different options. For example, a first-generation agent-based modeling analysis of urban infrastructure would include the following:

- (a) A determination of the spatial and temporal demand for urban infrastructure due to development/redevelopment activities
- (b) A listing of infrastructure alternatives that satisfy the needs of the development
- (c) A calculation of the material and energy required for the different alternatives
- (d) A determination of the vulnerability and resiliency of the urban infrastructure alternatives against exogenous or endogenous risks
- (e) An assessment of the local, regional, and global impacts of the alternatives through life cycle analysis

A second-generation model would include feedback loops in which all previous decisions are reconsidered in light of outcomes revealed by the model. This is more realistic because of future decisions do depend on evolving properties. Once the model is verified, the dynamic policies and economic drivers (i.e., market forces) that will yield more sustainable and resilient urban infrastructure can be determined more accurately. More importantly, the model can be used to assess the sensitivity of outcomes to different inputs and resources can be focused on those components that will result in the greatest improvement.

Methodologies

There are many methods for evaluating sustainability. As with all methods, each has strengths and limitations. Some of the methods used to evaluate the sustainability of urban environments are discussed below.

Resiliency

Resiliency is the ability of a structure to maintain its functionality despite exogenous or endogenous stressors. Resilient systems are flexible and adaptable and typically provide a better return on material investments in the long term. There are four important attributes that compose resiliency: robustness, redundancy, resourcefulness, and rapidity. Robustness is the ability of the system to withstand a given level of stress or demand. Redundancy is a measure of the inherent substitutability, or the ease with which a component or a whole system, or the function that they provide, could be replaced in case of a failure. Resourcefulness is a measure of the capacity to mobilize resources in the event of disruption. Rapidity is a measure of the capacity to contain losses or prevent further degradation in a timely manner. These 4 "R's" can be used as metrics in determining the sustainability of a system or infrastructure [7]. Urban infrastructure that is designed to minimize material and energy investments may not in fact be sustainable in the long term if the reductions result in low resiliency and the functioning of the infrastructure is jeopardized. Resiliency serves as a check against short-sighted, but well intentioned, actions that do not contribute to longer-term sustainability.

Life Cycle Assessment

Life cycle assessment (LCA) is the analysis of the impact of one particular product across all phases of the product's "life." Life can be defined in multiple ways. The most common way an LCA is performed is from cradle-to-grave. This considers raw material extraction, manufacturing, use, and disposal. While a great deal of attention is often paid to a product's manufacturing phase, the use phase is often the most important phase based on a product's impact. For example, a car uses a great deal more energy and resources in the use phase than in the manufacturing phase.

If a product is recycled or reused instead of being disposed of, it is cradle-to-cradle, and the life of the product is extended. For example, plastic soda bottles made of polyethylene terephthalate (PET, recycling symbol number 1) can be recycled to make new plastic bottles or a variety of other products, like T-shirts and carpeting. Cradle-to-cradle aims to reduce waste and decrease the amount of virgin resources that are used in a product.

Other methods may shorten the view of an LCA, such as cradle-to-gate and gate-to-gate. Cradle-to-gate looks at the product from raw resources all the way through manufacture, but ends there. This might be useful particularly to a manufacturing plant or for products that have negligible effect in their use and disposal phase. Manufacturing also might look at a gate-to-gate LCA. This kind of LCA looks only at the production chain. This could be useful in comparing the efficiencies of different manufacturing methods and tools.

Two additional LCA methods that may be employed are the Economic Input-Output LCA (EIO-LCA) and the Ecologically based LCA (Eco-LCA). The EIO-LCA was developed at Carnegie Mellon, and is used to measure the economic effects of the materials and energy used to make a product [8]. Ohio State University developed the Eco-LCA [9]. This tool adds to the traditional cradle-to-grave LCA by emphasizing more analysis on the direct and indirect effects of a product on pollutant emissions and economic flows.

Life cycle assessment is a useful tool when looking at sustainability; however, it does have a number of limitations. Not all factors can be tracked or easily given a numerical value. For example, it is difficult to

put a numerical value on the health effects of a product because toxicity and the transport and exposure routes may not be that easy to quantify.

One other limitation of LCA is the cutoff problem. For any product or process there are a large number of product stages or processes upstream and it is never possible to include all of them. Consider, for example, automobile manufacturing. An automobile needs parts for its manufacturing. The parts need a suite of metals. Mining equipment is required to obtain the required metals. But the metal needed for the mining equipment also required machines to obtain it. And so the chain continues. In LCA, the chain needs to be truncated somewhere resulting in errors of omission. Additionally, the different LCA methods and commercially available software, like SimaPro, that are widely used to conduct LCA may use different assumptions and data resulting in a variation of results that must be reconciled.

Material Flow Analysis

A material flow analysis (MFA) is a process used to assess the flow of mass and energy within a defined system. It can be conducted at national scales, at the scale of a product, or at any scale in between. Urban metabolism is a specific MFA tool looking at a defined urban area. This method tends to look at the city as a "black box." It considers the flows moving in or coming out of the city, but does not necessarily look at the mechanisms inside the city that cause these fluxes to occur. It does give a view of the amount of goods and energy needed to support a city, and how much waste the city produces. However, one constraint is the need for extensive input and output information for a huge number of materials. Some of these numbers are tracked at the national level, or are easy to track within one industry, but may be difficult to find or track for less controlled regions like states or cities. This approach works well for islands because all inputs are brought in by boats and recorded, and outputs also must go out on boat or stay on the island. With better data collection, this method will become increasingly useful. An improvement on urban metabolism would be to track the inflows of resources, the outflows of wastes, and the creation of infrastructure that creates wealth and comfort. The guiding principles would be to use fewer resources to create and operate the

infrastructure needed per capita. In practice, it can be difficult to compare one city to another because cities can be in different states of build-out of its infrastructure (e.g., a greenfield city as compared to a more mature city) and its climate, demographics, and topology can impose significant confounding effects.

Industrial Ecology

Industrial ecology can be defined as a multidisciplinary system-level study focusing on the interconnections between the industry, economics, and the natural environment. Industrial ecology utilizes a suite of different tools to analyze these interdependencies. Among these tools are life cycle assessment and material flow analysis, along with designing and manufacturing for the environment and ecological efficiency within the industry. The waste or by-products of one industry might be a resource for another industry, and if the two industries are located close enough for logistical viability, then the waste/by-product can be utilized by the other industry. For example, a cement manufacturer located near a power production facility may be able to use residual fly-ash from the power plant for its cement mix. And if a drywall production facility is also in the vicinity, it might be able to use the CaSO_4 from the cement plant as its raw material. Thus the net waste generation from these industries is reduced and their proximity saves significant fuel from the reduced need for transportation. This multitude of tools used in an industrial ecology assessment allows an industry or city to look at various designs and see how they affect inputs, outputs, and the overall effect on the economy.

Urban Sustainability

Urban sustainability is a systems-level holistic analysis of the sustainability of an urban system, particularly in regard to its infrastructure. Urban infrastructure is an intricate network of six major components: (1) private residences, (2) commercial establishments (including churches, schools, etc.), (3) water and wastewater, (4) energy, (5) transportation, and (6) land use. Urban sustainability also includes socioeconomic impacts and the policy drivers that cause the emergent properties of urban systems. With the demand for basic infrastructure growing with population, cities need to plan their infrastructures accordingly.

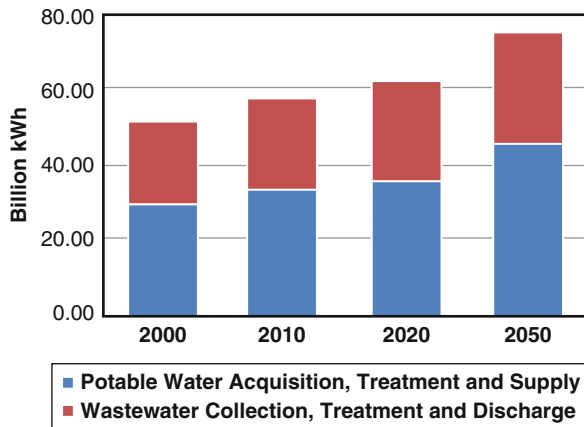
Traditionally, urban areas were built with the “big-pipe concept,” where the natural ecology was replaced with *hardscape* to provide for the urban infrastructure. *Nature* was restricted to curbside trees, and manicured parks and lawns filled with nonnative plants which require significantly more water and nutrients to thrive than the indigenous varieties. More recently, however, the importance of ecosystem services and preservation of natural ecology has been realized and natural alternatives, such as blue-belts for stormwater management, are increasingly being used in place of hardscapes. With this method, models are being developed to evaluate different options and to determine what social and economic decisions are needed to make them happen.

Current Infrastructure

Although current infrastructure may not be optimal, a large amount of capital has been invested in it and people still depend on it. One option for increasing sustainability is to develop a hybrid system of centralized and decentralized infrastructure, which increases redundancy without increasing the amount of requisite resources. Future planning also should guide how systems interact, and plan accordingly to save resources. Such interactions will now be described.

Water and Energy Nexus

Potable drinking water treatment provides clean drinking water, and wastewater treatment sanitizes sewage. Both use a great deal of energy in acquisition, treatment, and distribution. In 2000, the United States used more than 50 billion kWh of energy to attain, treat, and distribute drinking water and wastewater [10]. This is a huge demand for energy that is expected to continue to grow through 2050. The Electric Power Research Institute’s (EPRI’s) report on *Water and Sustainability* expects that by the year 2050, this number will reach over 75 billion kWh (Fig. 1). Already, water accounts for about 4% of the total energy sold in the United States. In California, it represents 18% of energy needs. And as water demand increases and good quality raw water becomes more difficult to find, these shares are likely to rise as advanced treatment technologies for reclaiming water or desalinating water are much more energy intensive. With this in mind, when a city is



Energy and Water Interdependence, and Their Implications for Urban Areas. Figure 1

Energy needed for drinking water and wastewater processes in the United States with data from the year 2000 and predictions of use until 2050 [10]

considering their water infrastructure, it also may want to keep in mind the amount of energy needed for various treatments for water and wastewater.

French Nuclear Power

In 2003, France experienced what was then their hottest summer on record [15]. The higher temperatures increased the demand for air-conditioning and refrigeration, which translated into an increased demand for electricity. In turn, increased energy production created more waste heat, which required more water for cooling purposes. Though France had plenty of capacity to produce electricity – the nation obtains ~80% of its electricity from nuclear power – the heat wave and drought limited the amount of water available for operation of the nuclear power plants, and the country was forced to turn off or reduce production at 17 of its nuclear reactors. As a result, France could no longer export electricity, which had been sold at €95 per MWh to other EU countries, but needed instead to import electricity at a cost of ~€1,000 per MWh.

Water can be treated using different methods, and with methodology differences there also are differences in energy dependency. Surface water tends to be

Energy and Water Interdependence, and Their Implications for Urban Areas. Table 1 National average energy demand by various water treatment systems [10, 11]

Water treatment and collection	kWh/MGal
Surface water treatment [11]	1,422
Groundwater treatment [10]	1,824
Brackish groundwater desalination [11]	3,900–9,750
Seawater desalination [11]	9,780–16,500

Energy and Water Interdependence, and Their Implications for Urban Areas. Table 2 National average energy demand by various wastewater treatment systems [10]

Wastewater treatment and collection	kWh/MGal
Trickling filter	955
Activated sludge	1,322
Advanced treatment without nitrification	1,541
Advanced treatment with nitrification	1,911

cleanest, and as such requires the least amount of energy to collect and treat (Table 1). Surface water is not available in all areas though. Groundwater requires slightly more energy to treat and seawater desalination can use up to 75 times the amount of energy as surface water.

For treating wastewater, methods include a trickling filter, activated sludge, advanced treatment without nitrification, and advanced treatment with nitrification. The energy demand, shown in Table 2, is lowest for the trickling filter and highest for advanced treatment with nitrification. The simplest treatments consume the least amount of energy, but they may not always be the most viable option given the quality of the water being treated and the requirements for the processed water. The gap in energy demand for the various approaches is much smaller in the case of wastewater than it is for water treatment.

Finally, power is also needed to distribute water. Logically, the farther that water needs to travel and the more difficult the water is to access, the more the energy that will be required to move it. Distributing water from lakes and rivers is the least energy intensive. It requires about 1,400 kWh/MGal to deliver [12].

Not all cities have lakes and rivers available to tap, however. For them, groundwater is more difficult to attain and requires slightly more energy. Lastly, discharging wastewater requires between 2,350 and 3,300 kWh/MGal on average.

The converse of energy being needed to provide, treat, and distribute water is that water is needed to produce, convert, and store energy. Water use in energy production varies depending on the source of energy. In the case of thermoelectric power generation, water is required for cooling purposes, while in case of hydro-electricity, water is used as the source of power. Water is also required for petroleum fuel processing, the primary energy source for transportation. Thus all forms of energy require water, which varies in the amount and purpose of its use depending on the energy source. Thermoelectric power accounts for 39% of freshwater withdrawals and 52% of our fresh surface water withdrawals [13].

The consumptive use of water for power generation is from evaporative losses. Every day thermoelectric power loses an average of 3.3 billion gallons of water to evaporation [13]. As one may suspect, certain kinds of power generation consume more water than others (Table 3). Accounting for the amount of energy produced and the water loss associated with each source, approximately 2 gal of water per kilowatt hour is lost to evaporation in the United States. Hydropower has an evaporative loss of 18.27 gal per kilowatt hour, which is more than 35 times the evaporative loss of water from coal-based thermal power. This implies that hydropower is feasible only in areas with an abundance of water. In contrast, photovoltaics and wind consume much less water.

Energy and Water Interdependence, and Their Implications for Urban Areas. Table 3 The evaporative loss of water by different energy sources in the United States [14]

Energy source	Gal/kWh (evaporative loss)
Hydro	18.27
Nuclear	0.62
Coal	0.49
Oil	0.43
PV solar	0.030
Wind	0.001

Example: Water and Energy Demand: A Tale of Two Cities

Atlanta and Phoenix are two cities with rapidly growing populations. As such, the demand on water and energy are rapidly growing. Being that they are in different regions, they have different demands for water. Table 4 describes how these cities are very different in terms of demand. For example, Atlanta uses more water indoors, but Phoenix uses a significant amount more outdoors. With long commutes and poor public transportation, Atlanta uses significantly more fuel per person per day than Phoenix. Some of the more important differences are due largely to Phoenix being a city in an arid region, very hot and dry, whereas Atlanta is more humid and wet generally. The water consumption for electricity production in Phoenix is more than four times that of Georgia. The electricity consumption for water supply and treatment also is around five times more in Phoenix than Georgia. With it being considerably more difficult to get water to Phoenix, these numbers are bound to become higher. The dependence of water on energy and energy on water in Phoenix are even greater than in Atlanta. Yet Atlanta has experienced a great deal of turmoil with its water resources, as described in the previous paragraph. So a strain on the Colorado River is subject to cause more damage, all things being the same. This table also shows that cities are not the same. No solution is right for all cities.

Beyond the Water and Energy Nexus

While water and energy are inherently connected, there are other connections that are equally important in the context of urban infrastructure (Fig. 2). Transportation needs energy to fuel vehicles, and the amount of energy available affects modes of transportation. Land use affects the amount of transportation activity in an area, and transportation is needed to made land accessible. The availability of water affects how land is used, and the way land is used affects the demand on water and the ways water can be treated. Likewise, the availability of energy affects land and water use, and the way the land is used affects the energy demands and the kind of energy that can be created. Additionally, water

Energy and Water Interdependence, and Their Implications for Urban Areas. Table 4 The relationship between water use and electricity use for the cities of Phoenix and Atlanta

		The City of Phoenix		The City of Atlanta	
Residential water demand	Indoor (gpcd)	48 ^a	1999 data in Chap. 4 of The Water Environment of Cities [37]	71	2001 data in The World's Water 2008–2009: The Biennial Report on Freshwater Resources
	Outdoor (gpcd)	110 ^a		20	
Power use	Residential electricity, kWh/person-day	36	2005 data [38]	41	2005 data Brown et al. (2008)
	Fuel, kWh/person-day	5.5		12.3	
Water consumption for electricity production	gal/kWh	7.85 (Arizona)	National Renewable Energy Laboratory (2003)	1.65 (Georgia)	Torcellini et al. (2003)
Electricity consumption for water supply and treatment	Water, kWh/MG	8,600	[39] Conveyance, 4600 Pumping, 2600 Water treatment, 100 Water distribution, 1300	1,700 ^b	City of Atlanta Watershed Management, [13, 36]
	Wastewater, kWh/MG	10,700	[39] Wastewater collection and treatment, 1500 Reclaimed water, 9200	1,830 ^b	

These data demonstrate the difference in use for the two different regions. Information of this sort must be considered when making energy and water plans for each city

^aThe numbers are modified assuming four people for one household

^bThe numbers are estimated based on the water and wastewater production of the City of Atlanta, the electricity use of the Atlanta Watershed Management Department (Thomas 2007), and the electricity demand for water supply and wastewater treatment of the South Atlantic Region (DOE [13])

is also used for extracting fuels and growing biofuels, which affects the transportation sector.

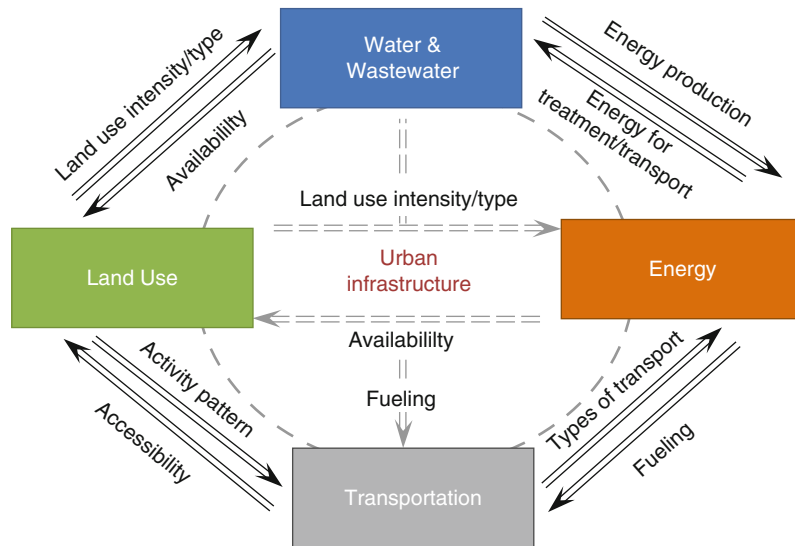
Energy for Transportation

Energy is needed to power a variety of transport vehicles including airplanes, trains, boats, trucks, and automobiles. The fuel used in these vehicles may be equally diverse and may include gasoline, diesel (both petro and bio based), ethanol, methanol, natural gas, electricity, hydrogen, solar, and even coal and wood (though these latter two are nearly negligible in the modern vehicle mix). The latest trend is to create hybrid vehicles that can utilize two or more fuel sources (e.g., gasoline and electricity). As of 2009, the transportation sector accounts for ~28% of the of total

energy consumption in the United States and 72% of the nation's petroleum consumption [16]. Transportation also accounted for 3% of natural gas consumption and 12% of renewable energy use, mostly in the form of ethanol. Given these significant shares, the effect of transportation on energy needs to be considered when planning cities or city improvements.

Transportation, Land Use, and Energy

As the need for mobility increases, improving existing vehicle efficiency will lead to marginal reductions in the environmental impacts of transportation. Greater improvements, however, may be possible if switching to other modes of transportation is considered. In a 2007 study of plug-in hybrid electric vehicles, the



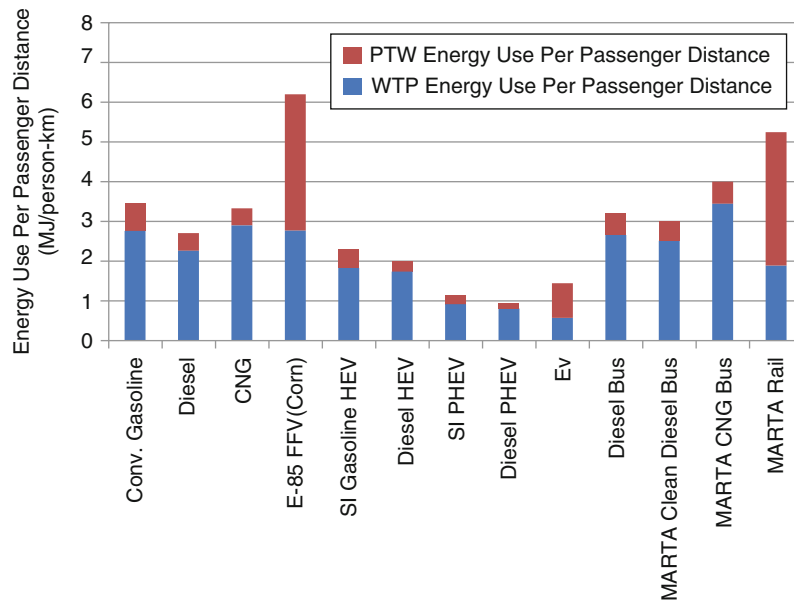
Energy and Water Interdependence, and Their Implications for Urban Areas. Figure 2
Water, energy, transportation, and land-use systems are interconnected in urban regions

Pacific Northwest National Lab [6] found that the current US electric power infrastructure could support the energy needs of 43–73% of the current light-duty vehicle fleet – the lower range resulting from the limiting of vehicle charging only to the 6 PM to 6 AM period. The study further estimated that if 73% of the US fleet was converted to electric vehicles, the United States would no longer need to import oil, and greenhouse emissions could be reduced by 27% (even if all electricity were generated by coal-fired power plants). While such broad-scale analyses are informative, it is important to note that complex local relationships underlie these bigger developments.

Bras recently completed an assessment of the energy and carbon dioxide emissions from different modes of transportation in Atlanta. Many kinds of vehicles and fuels were included in the study: conventional gasoline, diesel, compressed natural gas (CNG), ethanol 85 flexible-fuel vehicles (E85 FFV), spark-ignition (SI) gasoline hybrid electric vehicle (HEV), diesel HEV, SI plug-in hybrid electric vehicle (PHEV), diesel PHEV, electric vehicle, diesel vehicle, MARTA Clean Diesel bus, MARTA CNG bus, and MARTA rail. (MARTA is the Metropolitan Atlanta Rapid Transit Authority, which is the primary form of public transportation

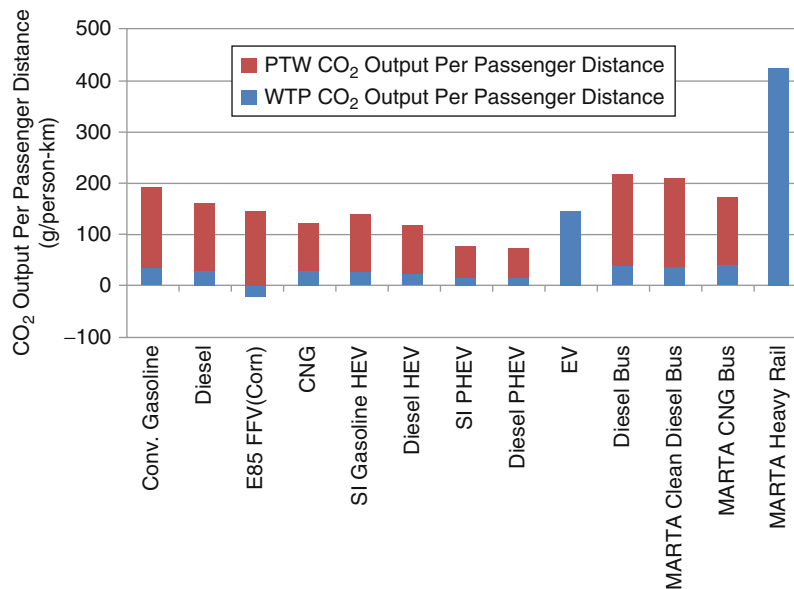
in Atlanta.) All of these transportation methods were evaluated on a Well-To-PUMP (WTP) and Pump-To-Wheels (PTW) basis for energy use and carbon dioxide emissions (see Figs. 3 and 4). Due to low ridership, MARTA's buses and rail are among the worst for both carbon dioxide emissions per passenger distance and for energy use per passenger distance. Increasing ridership would greatly improve these metrics. Other results suggest that electric vehicles do not fare much better than conventional gasoline and diesel. This is due, in part, to Georgia's dependence on older inefficient coal-fired power plants for energy. Overall, plug-in hybrid electric vehicles used the least energy and emitted the least carbon dioxide per passenger distance.

The implications of the effect of ridership on energy use and carbon dioxide emissions from various transportation modes in Atlanta suggest that other variables beyond just vehicles and fuel are also important. Planning and developing cities for growth with compact land use and transportation could significantly improve their sustainability and lower their dependence on energy. The way land is used affects the modes of transportation available and their efficiencies. For example, a compact city favors walking,



Energy and Water Interdependence, and Their Implications for Urban Areas. Figure 3

The energy use per passenger distance for various transportation modes in Atlanta, GA



Energy and Water Interdependence, and Their Implications for Urban Areas. Figure 4

Carbon dioxide emissions per passenger distance for various transportation modes in Atlanta, GA

heavy rail, and other modes of efficient public transportation. New York City is a good example of this kind of city, where parking fees, gas expenses, and traffic lead the majority of the population to walk,

take the subway, or ride the bus. An opposing example might be Atlanta, where the city is not compact enough to offer many people the option to walk or use public transportation.

Water for Transportation

Along with energy and land use, water is also intertwined with transportation. Department of Energy data [13], along with data from Harto’s LCA of low carbon fuels [17], shows that oil consumes between 0.03 and 0.08 gal of water per kilowatt hour of energy [13] (see Table 5). Natural gas production uses a similar amount. Water needed to produce biodiesel from irrigated soy, though, is in the range of 0.39–9 gal of water per kilowatt hour. Corn-ethanol needs up to 19 gal of water per kilowatt hour, or ~500 times more than that for oil. From another perspective, Webber [12] found that a traditional gasoline car consumes around 7–14 gal of water per 100 miles traveled, PHEVs consume ~24 gal per 100 miles, and ethanol fueled vehicles require 130–6,200 gal per 100 miles.

The Compounding Power of Density

Urban density reaps a great many sustainable benefits. Public transportation is more accessible and commuting distances are generally shorter. Compact living also means that buildings can be closer to the

infrastructures that provide them water and energy, which means less energy and materials are needed for distribution. It is also often the case that the more complex power and water networks that serve dense urban areas have built in redundancies that can compensate for component failures within the system. Finally, density affords other cross-pollinating opportunities such as combined heat and power, and low-impact development water infrastructure that also provides green space for recreation.

Jobs, Quality of Life, and Tax Revenue

The nexus of infrastructure extends beyond the interconnection of the physical, functional, and environmental constituent pieces. It also has implications for a place’s social and economic well-being. Consider again, for example, the benefits of compact urban design. Land consumed by large private residential lots cannot be used by the collective community. Small private lots, on the other hand, preserve the opportunity to leave larger tracts available for recreation, wildlife habitat, and other amenities that are not possible when the land is fragmented. These public conveniences and desired attributes improve the quality of life for everyone, create a more favorable, stable, and job-producing business climate, and raise property values. And when the resultant increase in tax revenues is reinvested in the infrastructure, an engine of prosperity is created (see Fig. 5).

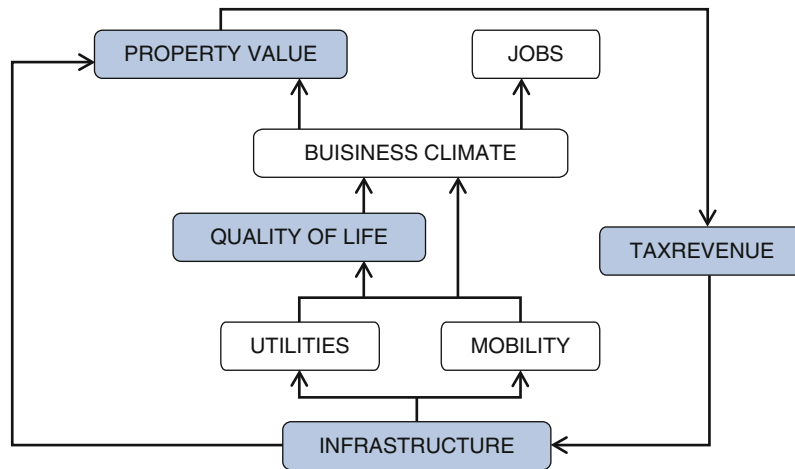
Energy and Water Interdependence, and Their Implications for Urban Areas. Table 5 The dependence of various transportation fuels on water

Unit: Gal/kWh	Low	High	
Coal [13]	0.007	0.027	Mining+washing
Petroleum/oil [13]	0.03	0.076	Extraction+refining
Natural gas [13]	0.01	0.01	Extraction+processing
Corn-ethanol [17]	1.26	19	Assuming 15% irrigation for USA
Cellulosic ethanol [17]	0.13	0.431	No irrigation
Cellulosic ethanol [17]	16	19	Irrigation
Soy-biodiesel [17]	0.392	8.98	Assuming 4% irrigation for USA
Algae biodiesel [17]	0.839	1.762	Enclosed
Algae biodiesel [17]	0.895	18.351	Open

Steps Toward a More Sustainable Future

Water

Rain water is a source of clean water that does not need much treatment. However, most cities currently do not handle this resource efficiently. In most cases, stormwater becomes contaminated when it falls onto urban or agricultural surfaces, or is shunted into the wastewater system. Collection and treatment of stormwater using low-impact development (LID) techniques offers a myriad of benefits. The first step is to separate the collection of stormwater from wastewater. One immediate benefit is that with less water to be treated, less energy is needed in the water treatment process. Reduced flow into the wastewater stream also reduces the risk of sewage overflow into surface waters.



Energy and Water Interdependence, and Their Implications for Urban Areas. Figure 5

The relationship between improved infrastructure and an improved socioeconomic environment

The stormwater that is collected can be used to create green spaces, to flush toilets, for cooling, or for fire fighting. Below is a sample of LID techniques that can be used.

Vancouver

The City of Vancouver was concerned about the stormwater and sewage that was being discharged into its salmon-bearing rivers [18]. The city considered traditional separation and decentralized treatment, but after discovering that putting in a centralized stormwater treatment facility would cost the city an estimated \$4 billion, it decided to look into different options. One approach that employed low-impact development and decentralized treatment satisfied the water treatment needs while also creating green space which increased property values and led to an estimated increase of \$400 million in tax revenue. It also resolved the environmental problem for their salmon-bearing rivers, improved conditions for other wildlife, and reduced the urban heat-island effect.

herbicides, nutrients, other existing pollutants, and emerging contaminants like pharmaceuticals and personal care products from moving from the land to the water. Some of these techniques may simply reduce the amount of toxins used in a process or pesticides used for farming. Some structural techniques include detention basins, which are used to prevent flooding in storms, and pervious pavement, which allows water to flow through its pores while holding a number of pollutants in the soil. These techniques help manage the flow of stormwater so that wastewater systems are less likely to be overwhelmed. They often rely on natural techniques, such as bioremediation or wet land treatment, to clean the water and are often less expensive and less energy intensive alternatives.

New York

The Safe Drinking Water Act required New York City to build new water filtration plants at a cost of ~\$8 billion. Balking at the cost, the city instead acquired undeveloped lands to buffer the watershed, initiated a host of Best Management Practices (BMPs), worked with the farming community to develop a Whole Farm Plan to reduce runoff pollution, used LID techniques to reduce stormwater flow, and made a few upgrades to existing sewer and septic systems in the region. In the end, these actions, in lieu of the \$8 billion cost of new

Best Management Practices for Pollution Control

Best Management Practices (BMPs) are techniques used in controlling water pollution and the flow of water. They are put in place to control pesticides,

filtration plants, enhanced the economic productivity of the land from recreation, farming, and the protection of the watershed from runoff [19].

Green Roofs Green roofs are roofs that are partially or wholly covered with vegetation. They increase urban green space, reduce stormwater runoff, sequester carbon dioxide directly, and reduce building energy consumption. They can be added to buildings in two ways, intensive or extensive. An intensive roof is normally between 8 and 12 in. thick and weighs 80–120 lb per square foot [20]. Extensive roofs are thinner and much lighter (10–50 lb per square foot). Intensive roofs are often built as gardens, but must be well supported due to the increased weight. Extensive roofs are not necessarily meant to be visited, but may be used for some of the same benefits, but at a lower cost.

Indigenous Plants Plants that are native to a region require less water, fertilizers, and pesticides than exotic plants. Indigenous plants have adapted to local conditions. Less effort is needed to maintain these plants, and they live longer than other plants would in their environment. Alternatively, plants with superior carbon sequestration qualities like yellow poplar and American sweet gums more effectively reduce the carbon footprint of a city relative to slower growing hardwood trees.

Pervious Pavement Pervious pavement allows water to pass through the pavement for groundwater recharge and thereby decreases the amount of rainfall that enters the wastewater system. Pervious pavement is ideal for large, low traffic uses like parking lots, driveways, bike paths, and sidewalks. Not only does pervious pavement reduce the load of stormwater on wastewater treatment systems, but it also reduces nutrient and metal loads from runoff (by 80% and 90%, respectively) [21].

LID techniques should be accessed city by city and based on a cost-to-benefit analysis, as not all techniques are proper for each city circumstance. But while most LID techniques are beneficial even within traditional centralized water systems, decentralized

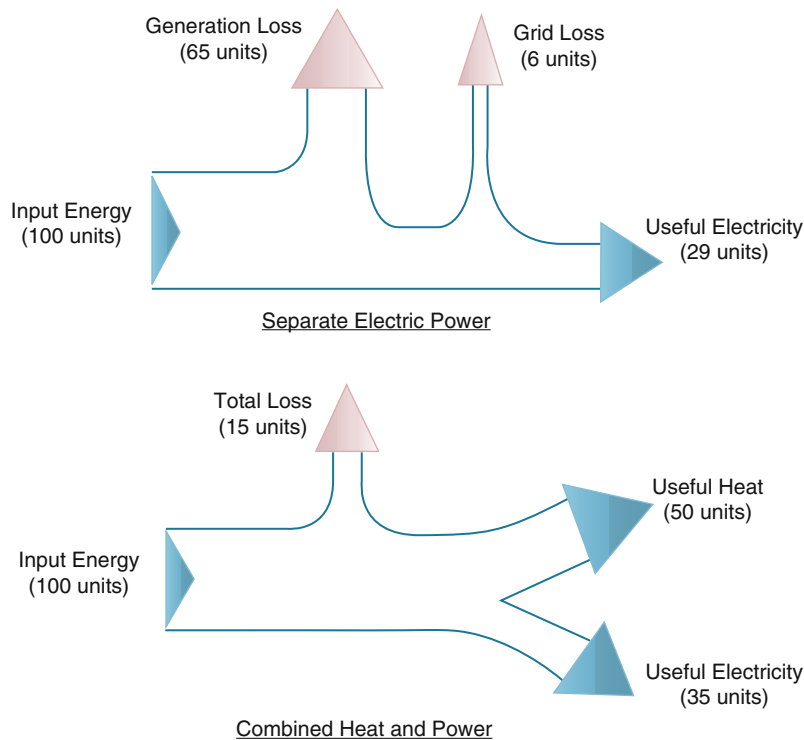
systems may offer additional benefits. With conveyance and distribution responsible for 80% of the energy demand in the water sector [10], decentralized systems can offer significant savings. In addition, due to the lower residence time of the water in the distribution system, the final water quality at the tap is less compromised. Decentralized systems also add resilience to the system due to higher redundancies than in centralized systems. That is, if a centralized system fails, all the customers that it serves are at risk of being disconnected from their water supply. When a decentralized system fails, another nearby station can supply the water needed until the damaged plant is fixed.

While newly developing areas may implement decentralized water systems, other cities with existing centralized water treatment systems should not look to totally disassemble their infrastructure. Instead they should consider developing a hybrid centralized/decentralized model. This might be as simple as adding a decentralized plant in a quickly growing area instead of expanding the centralized plant. And because LID techniques such as rainwater harvesting and low flow toilets reduce the output needed from a water treatment plant, LID techniques and decentralization simultaneously work to enable and enhance each other.

Energy

Fossil fuels are a limited resource. As such, their use needs to be judicious and efficient. Improving the efficiency of vehicles, lighting, appliances, electronics, heating, ventilation, and cooling is one strategy. Another strategy may be redesigning the energy system by distributing power and incorporating renewables, and truly integrating multiple resource planning. Below are described a sample of opportunities.

Combined Heat and Power A large amount of energy is lost in creating electric power; only about 30% of the energy input results in electricity delivered to the customer [23]. Seventy percent of the energy is lost in converting the primary fuel into electricity and distributing through the power grid. For that same amount of input energy, a combined heat and power



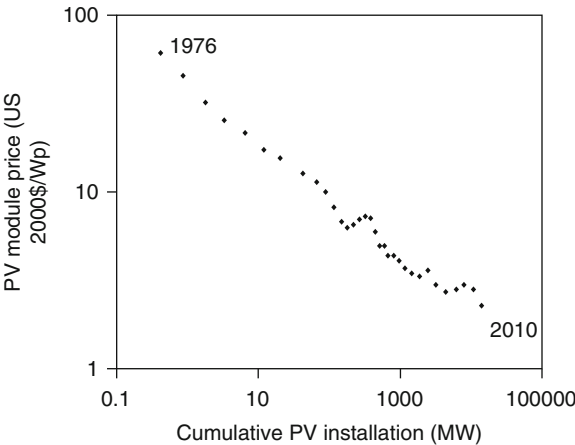
Energy and Water Interdependence, and Their Implications for Urban Areas. Figure 6

Utilization of energy for separate electric power and combined heat and power. The amount of energy lost is much higher in the separate system due to the recycling of heat to make more energy (Amended from [22])

system could utilize $\sim 85\%$ of the input energy (see Fig. 6). Combined heat and power has the potential to provide the United States 20% of the electricity by 2030, which could reduce an estimated 0.2 gigatons of carbon annually. This system also emits, on average, 1/10th of the nitrogen oxides (NO_x) per kWh as compared to the average grid electricity [24]. Helsinki Energy in Finland used combined heat and power, and has experienced great results [24]. Sulfur dioxide emissions from district heating were reduced around 90% from 1980 to 2004. This scheme also saved Finland six million tons of carbon emissions in 2004. Using less energy means fewer greenhouse gases are created, and with the climate changing so quickly, emission reduction is needed. Systems like these generate electricity and use the waste heat for a variety of processes like chemical manufacturing, heating buildings, or producing hot water, and thereby eliminate or

reduce the energy and consequent environmental impacts needed for those demands [23, 24].

Energy Efficient Products In most existing applications, simply upgrading or modernizing energy demanding products will result in improved efficiencies and higher performance. Even low-end refrigerators produced today may be several times more efficient than refrigerators produced 20, 30, and 40 years ago, while also providing more features and functionality along with a relatively short payback period. When choosing new or replacement products, however, certified Energy Star products use 20–30% less energy than the federal standards require [25]. Eligible products include appliances, electronics, heating and cooling systems, lighting, fans, plumbing fixtures, and building products like insulation and roofing.



Energy and Water Interdependence, and Their Implications for Urban Areas. Figure 7

Price of a PV module per Watt peak (constant year 2000 US dollars) as a function of cumulative installation of PV in megawatts for the period 1976–2010 [27]

Photovoltaics and Other Renewable Energy Sources

Every day the sun provides more than enough energy to meet all the demands of humans and nature. But while nature has evolved to efficiently harvest this free energy, human made systems are only now becoming cost-effective (see Fig. 7). Comparing photovoltaic power generation to coal generation, Table 6 shows for each source the ratio of energy output to energy invested water use, land use, cost of power, and jobs. While the lower overnight cost of power still favors coal, the other attributes indicate that PV is already a competitive alternative. With costs continuing to drop, however, PV could soon become the preferred power source with coal becoming the less attractive alternative.

Wind, geothermal, biomass, and hydroelectricity are some of the other renewable energies that are growing in importance. For example, a recent working report suggested that there is an unrealized potential of 29,400 MW of hydroelectricity that could be developed in the United States [26]. One problem with renewable resources, though, is their inherent intermittency. In order for many renewables to become practical, large-scale inexpensive energy storage methods need to be developed. These methods may

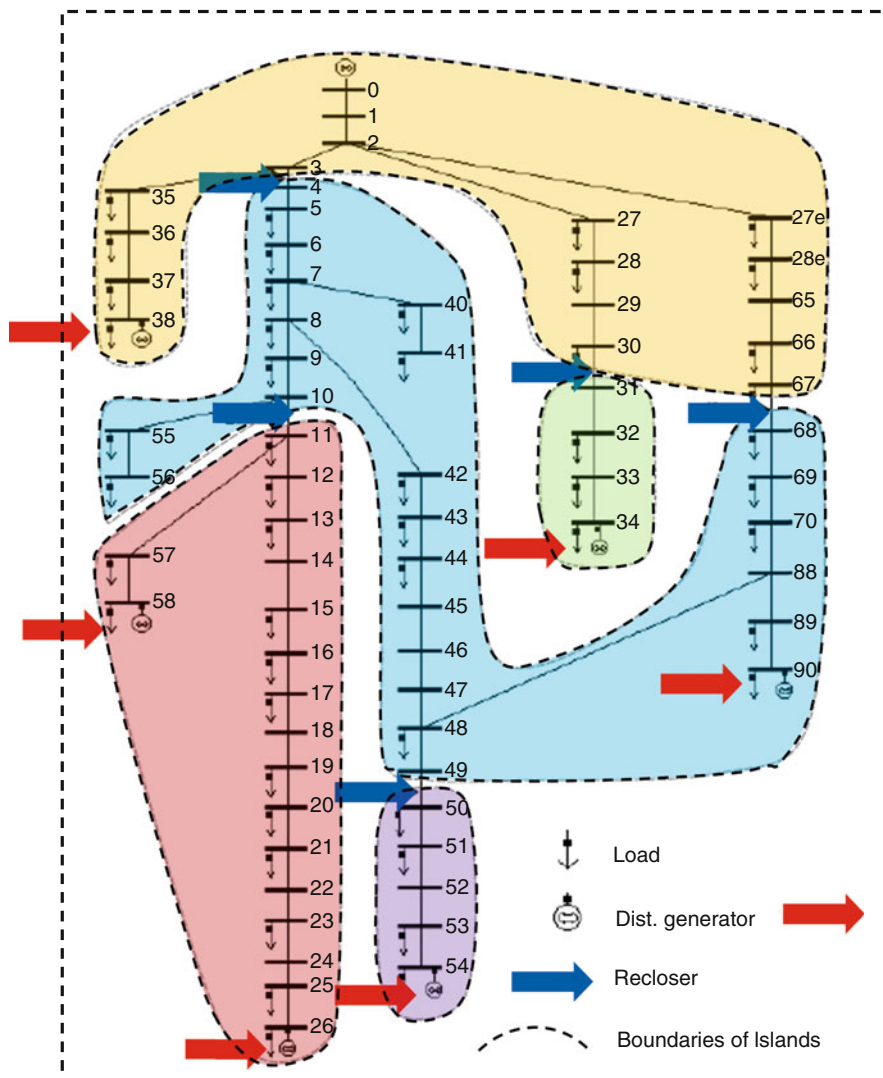
Energy and Water Interdependence, and Their Implications for Urban Areas. Table 6 A comparison of power generation from photovoltaic and coal energy sources

Energy Source	Photovoltaic	Coal
Energy output Energy invested [28]	9.0	5.0
Water use [29]	0.001 Gal/kWh	0.49 Gal/kWh
Land use	0.51 kWh/acre (RET Screen Simulation)	690 kWh/acre [30–33]
Overnight cost [34]	\$4.75/kWh	\$2.84/kWh
Total job-years/GW h (avg) [32]	0.87	0.11

include various types or combinations of electrochemical batteries, flywheels, compressed air, superconductive magnetic storage, and the oldest, the least capacity-constrained and most popular when available, pumped hydro-storage.

Decentralizing Distribution

Apart from the manner in which power is generated, system resiliency can be increased by changing how the power is distributed. The smaller scales inherent to renewable generators, such as PV, mean that these generators can more readily be located in various locations across the distribution feeder. Microgrids can be created by separating the distribution feeders into zones of balanced generation and load can by placing reclosers at the zone boundaries (reclosers are similar to circuit breakers, but do not have the capability to interrupt the fault current; see Fig. 8, e.g.). In the case of a feeder separated from one of its supply substations by a fault location, such a network may allow the feeder still to be energized and supplied by other generators while the faults are being mitigated. That would improve reliability as measured by indices such as SAIFI and SAIDI (system average interruption frequency and duration index), CAIFI and CAIDI (customer average interruption frequency and duration index), and ASUI and ASAI (average service (un)availability index). Improvements in such indices (or composite indices) usually mean better customer satisfaction and diminished losses associated with repetitive outages, that is, increased resiliency.



Energy and Water Interdependence, and Their Implications for Urban Areas. Figure 8

A depiction of an islanded distribution network. This system is split into multiple zones with balanced distributed generation and loads (as shown by the *dotted lines* and *multiple colors*). There are multiple distributed generators, and each island has a recloser at the boundaries with other zones. This allows the islands to cut off from each other in case of a disturbance, keeping the larger portion of the network supplied with energy. Once the problem is fixed, the islanded system can then automatically be reconnected to the grid (transmission network)

Example: Distributed Renewable PV Generation

Distributed generation consists of small-scale and decentralized electric energy systems. Capacities vary, typically in the range of several kW to hundreds of MW. Accurate and efficient system analysis algorithms are

needed in order to analyze the impact of PV systems on various types of microgrids and distribution networks. The influence of uncertainties can be modeled via suitably optimized Monte Carlo techniques.

Commercial PV power industries started developing in the 1970s. In spite of a 70% reduction of the real price

of PV modules over 40 years, energy from PV remains too expensive to compete with conventional sources. Political changes in the United States in the early 1980s ended substantial funding for solar energy research and, since the nation represented nearly 80% of the global market for solar energy at that time, virtually halted solar energy development around the world. In 2004, renewable energy sources accounted for 9.6% of the electricity generated in the United States and 18.6% worldwide. China was the leading nation by renewable generation capacity in 2008 (598 TW h of renewable energy produced that year). A total of 3,584 TW h of renewable energy was produced in the world in 2008, of that only 12 TW h was from solar photovoltaic generators while, for example, 16% (3,288 TW h) of world electric energy was produced from hydroelectric plants. However, PV has become one of the most rapidly growing energy generation technologies. PV module shipments have grown at an average annual rate of 40% since 1996, up from 13% in the previous decade. The potential growth may transform PV into a \$100 billion industry [35].

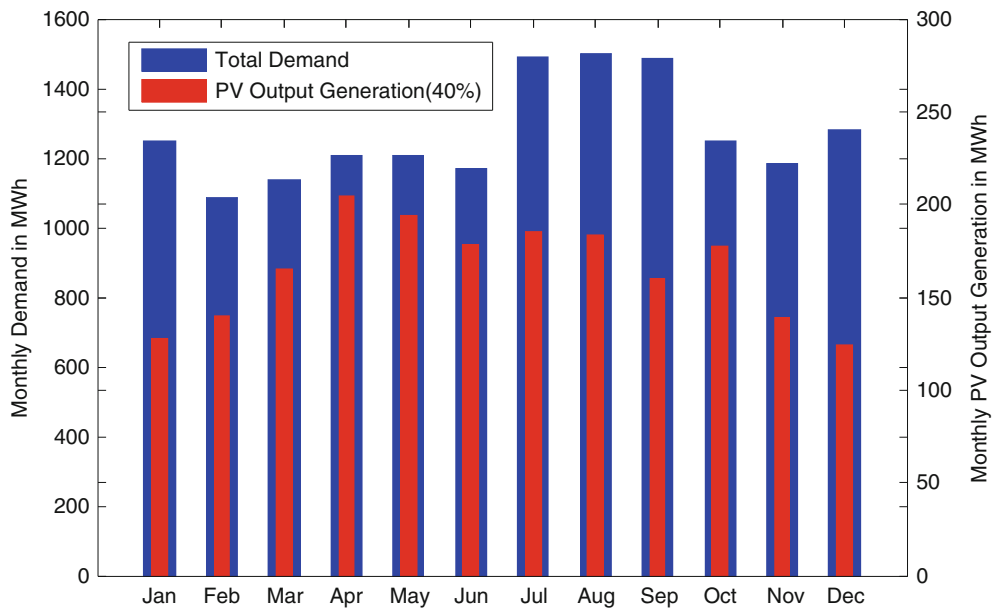
On the technical side of PV proliferation, electric power utilities are not motivated to allow interconnections of customer-owned generators (such as many PV installations would be) to their distribution networks (mostly for operational reasons). Utilities tend to put nonutility generation under the extensive technical analysis. Conversely, the regulating authorities tend to act in favor of DG owners and support that the interconnection be as easy and transparent as possible.

The main objective of PV installation is to boost the energy savings. Figure 9 shows annual estimated monthly PV generation output of an assumed PV system in the Atlanta area in the United States. Figure 10 presents how the total load demand is decreased due to the PV generation with capacity ranging 10–40% of the annual peak load at the same feeder. The figure represents the system performance on the 244th day of the year. If the distributed generator (DG) is owned by the customer, the actual benefit is cost saving in the electric bill. If the DG is owned by the utility, the benefit is avoided energy production from the less desirable or more expensive sources and reduction in transmission and distribution losses. This benefit means energy saving in both cases.

The second impact of solar electricity generation is the ecological impact. The PV output energy savings provide the opportunity to create the avoided carbon footprint by not using the more ecologically impactful technologies (coal, natural gas, etc.). In addition to reducing the effective feeder load in the distribution networks where it is installed (thus also reducing the transmission and distribution losses), PV generation can also bring about CO₂ reduction. It can also cause water footprint reduction through avoided thermoelectric and hydroelectric generation, and fuel reduction, such as coal, petroleum, and natural gas in the thermal power generation as in Table 6. The example assumes that the PV system is optimally oriented to maximize energy production. (If shaving of peak load levels is the main objective, it is possible to orient the PV system more westward, thereby boosting the PV output at the peak load times.)

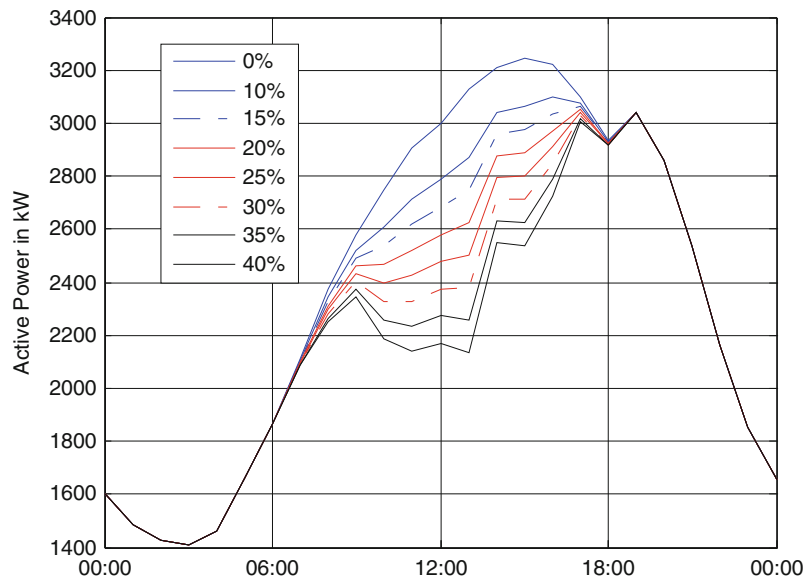
Understanding Aggregate Demand A number of factors need to be considered in every city when making decisions to improve sustainability. Inputting all the data necessary to run a good model of a city is a massive undertaking, but the model could be a very useful tool.

For the modeling of Atlanta, Geographic Information System (GIS) tools were used to input infrastructure data. The What if?™ system was used to model Atlanta. This software was used to show the suitability of various infrastructures for an area. Figure 11 shows the suitability for eight different land-use types. As Atlanta continues to develop, this can be used as a tool to plan infrastructures for areas that are more suitable. Along with suitability factors, this program was used to predict the land use of Atlanta as it grows. Figure 12 shows Atlanta's use of land in 2010 and the prediction for land use in 2030, using the Business as Usual scenario. This scenario looks to grow the city in the traditional urban sprawl way, and the figure shows much of the land use going to residential use and employment centers expanding outward from the city center. Figure 13 offers a different comparison, this one of two different growth scenarios: Business as Usual and Compact Growth as predicted for 2030. This comparison shows the difference in how Atlanta



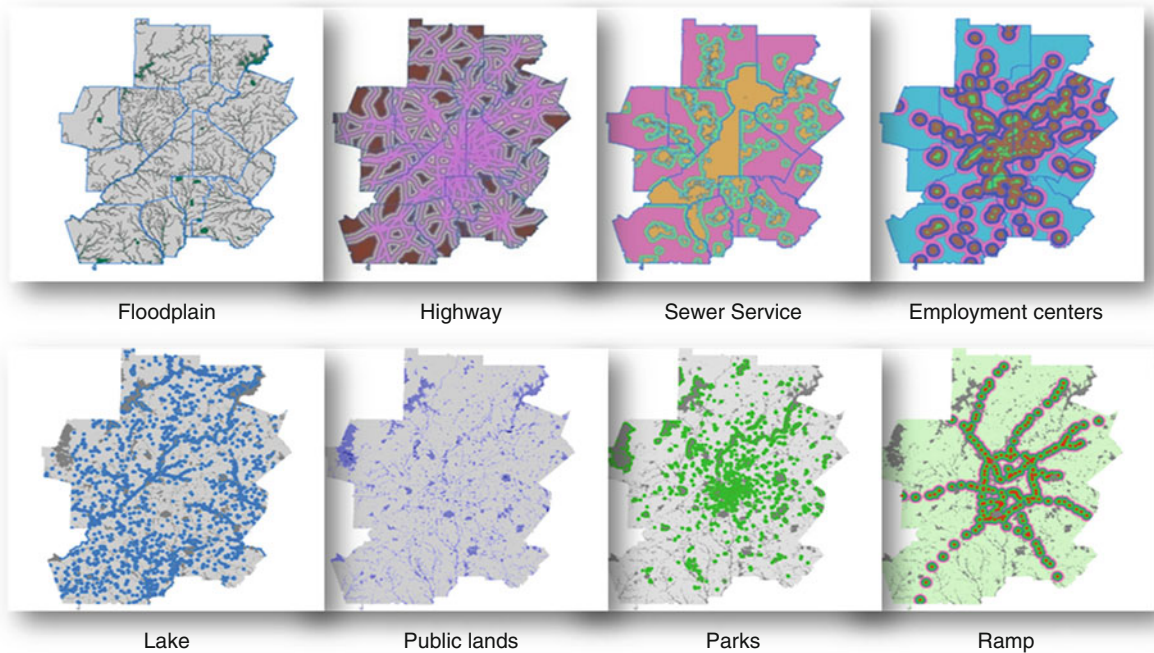
Energy and Water Interdependence, and Their Implications for Urban Areas. Figure 9

Annual PV generation and load demand on a typical distribution feeder. It is assumed that the total PV capacity is equal to 40% of the peak load demand

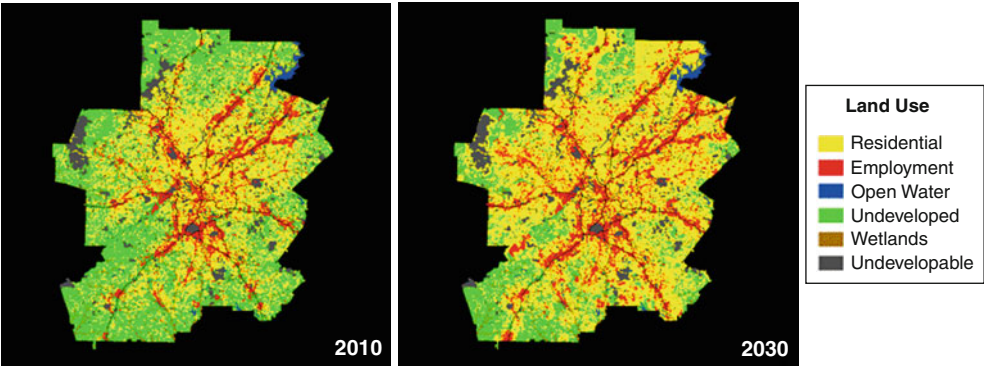


Energy and Water Interdependence, and Their Implications for Urban Areas. Figure 10

The effect of PV generation on a typical day of the year (horizontal axis represents time of the day). Assumed PV capacity is 40% of the annual peak load



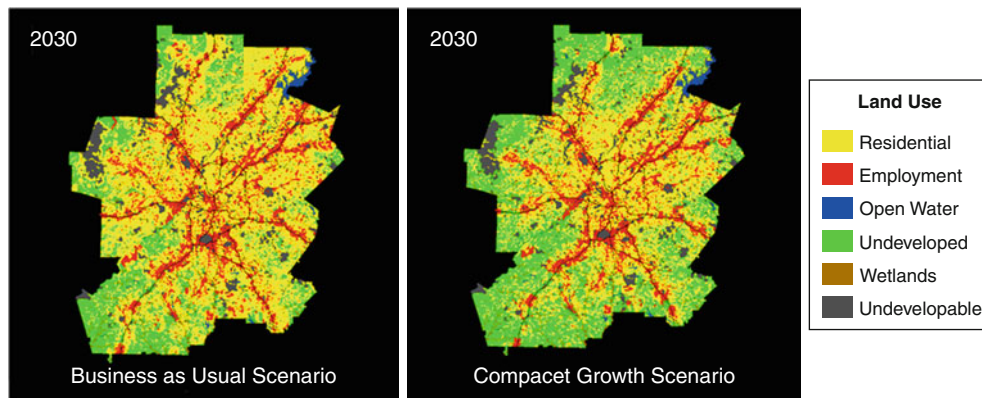
Energy and Water Interdependence, and Their Implications for Urban Areas. Figure 11
The maps show the 13 county metropolitan Atlanta area. Based on the land, the program What if?TM developed the suitability for different kinds of infrastructure and land use



Energy and Water Interdependence, and Their Implications for Urban Areas. Figure 12
The 13 county metropolitan Atlanta area's land use. This shows the prospective land use using the Business as Usual scenario from the What if?TM software. Employment centers and residential areas expand outward over the 20 years, leaving very little area undeveloped

can grow, with more compact living and more undeveloped area, or with continued urban sprawl. Programs such as What if?TM and UrbanSim give a rough prediction of the future, but the scenarios in these programs are not easily edited by the user. To date, infrastructure planning has been an exercise in

developing a few different scenarios that fit the general tendencies of the agents and groups of agents that the infrastructure intends to serve. This approach, though, fails to account for a number of factors that affect the decision making of agents. Looking beyond a set of scenarios, agent-based modeling creates



Energy and Water Interdependence, and Their Implications for Urban Areas. Figure 13

The 13 county metropolitan Atlanta area's prospective land use in 2030. This is according to the simulation run with the What if?TM software. These show the comparison of development based on two different scenarios, Business as Usual and Compact Growth. Compact Growth shows that much more land is left undeveloped and residences remain near the city, whereas Business as Usual shows an increase in urban sprawl

equations for how people react to different constraints and opportunities. For example, surveys and behavioral analyses can be used to develop algorithms that describe how people react to policy decisions or to gauge consumer awareness. These reactions though are tempered by the context of many other variables though such as economic feasibility, the state of current infrastructure, or the availability of technology. Modeling could provide a number of answers to questions concerning a city's sustainability. Some of the questions it could help address are as follows:

- What price are people willing to pay for renewable energies, and what kind of policies may be needed to encourage this choice?
- How can the transportation network be better planned in order to make public transportation efficient, and at what price will people pay to choose more sustainable forms of transportation?
- What incentives are needed for people to live in more compact spaces?
- How can consumption of nonrenewable resources be reduced?

Agent-based modeling, in combination with other modeling tools, is needed to predict the demand for urban infrastructure, how it should be built and designed, and what materials should be used.

Future Directions

For the future, there is a need to redesign our anthroposphere (the place where we live), so that it can exist within the means of nature: using only the renewable resources that nature can sustainably provide, while generating only wastes that nature can sustainably assimilate.

Future goals include:

- Monitor, model, visualize, and predict the emergent properties of urban infrastructure systems and their resilience to stressors.
- Understand the flow of resources and information as they move through the urban system with urban metabolism.
- Develop a pedagogy for the design of complex systems for sustainability within the context of urban systems.
- Integrate the human perspective into urban infrastructure so that socially sustainable policies and outcomes are produced.
- Research and develop sustainable alternative technologies for water treatment, and materials and energy production.

Acknowledgments

The authors would like to thank Brook Byers Institute for Sustainable Systems, High Tower Chair, and Georgia Research Alliance for their financial supports.

Funding for this research was provided by the US NSF grant 0836046: EFRI – RESIN Project: Sustainable Infrastructures for Energy and Water Supply (SINEWS). Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the view of the supporting organizations.

Bibliography

Primary Literature

- GreenBiz (2000) Why tackle the sustainability challenge? GreenBiz.com
- Committee on Education (1987) Our common future, from one earth to one world. In: UN Documents: gathering a body of global agreements, UNO
- Xu M, Crittenden J, Chen Y, Thomas V, Noonan D, Desroches R, Brown M, French S (2010) Gigaton problems need gigaton solutions. *Environ Sci Technol* 44:4037–4041
- Toepfer K (2005) One planet many people. In: UNEP press release, 2005/27
- UNEP (2009) Population division of the department of economic and social affairs of the UN Secretariat (2009). World population prospects: the 2009 revision and world urbanization prospects. United Nations Population Division, New York
- Kintner-Meyer M (2007) Impacts assessment of plug-in hybrid electric vehicles on electric utilities and regional US power grids part 1: technical analysis. Pacific Northwest National Laboratory, Richland
- Bruneau M, Reinhorn A (2007) Exploring the concept of seismic resilience for acute care facilities. *Earthq Spectra* 23:42–62
- Carnegie Mellon University Green Design Institute (2008) Economic input-output life cycle assessment. <http://www.eiolca.net>
- Ohio State University Process Systems Engineering Group (2009) Ecologically-based life cycle assessment. <http://resilience.eng.ohio-state.edu/eco-lca/index.htm>
- Goldstein R, Smith W (2002) Water and sustainability (vol 4): U.S. electricity consumption for water supply and treatment – the next half century. EPRI Technical Report 1006787, EPRI, Palo Alto
- Klein G, Krebs M, Hall V, O'Brien T, Blevins BB (2005) California's water-energy relationship. In: Integrated energy policy report proceeding (04-IEPR-01E). California Energy Commission, Sacramento
- Webber M (2008) Catch-22: water vs. energy. *Sci Am* 18:34–41
- US Department of Energy (2006) Energy demands on water resources. Report to Congress on the interdependencies of energy and water. US DOE, Washington, DC
- Gipe P (1995) Wind energy comes of age. Wiley, New York
- Kanter J (2007) Climate change puts energy into hot water. *New York Times*, 20 May 2007
- US Energy Information Administration (2010) Annual energy review 2009. US Department of Energy, Washington, DC
- Harto C, Meyers R, Williams E (2010) Life cycle water use of low-carbon transport fuels. *Energy Policy* 38:4933–4944
- Engineering Services, City of Vancouver (2005) Crown street, sustainable streetscapes and fish habitat enhancement project. Engineering Services, City of Vancouver, Vancouver
- USAID Water Team (2003) USAID water team case study in integrated water resources management. USAID, Washington, DC. http://www.usaid.gov/our_work/environment/water/case_studies_success.html
- Culligan P (2011) Solving urbanization challenges by design – the science of green roofs. Columbia University, The Earth Institute, New York
- Minnesota Stormwater Steering Committee (2008) The Minnesota stormwater manual: version 2. Minnesota Pollution Control Agency, St. Paul, pp 352–353
- EPA (2011) WaterSense®: program accomplishments. http://www.epa.gov/WaterSense/about_us/program_accomplishments.html
- Elliott RN, Spurr M (1999) Combined heat and power: capturing wasted energy. ACEEE, Washington, DC
- UNEP (2005) Combined heat and power: clean and efficient energy production. United Nations Environmental Programme, Governing Council, <http://www.unep.org/GC/GCSS-IX/Documents/FINLAND-combinedheatpower.pdf>
- US Energy Information Administration Database
- Weise E (2008) The shape of lights to come? Not everyone's buying it. *USA Today*, 28 Feb 2008
- Brown M, Gumerman E, Baek Y, Morris C, Wang Y (2010) Renewable energy in the south: a policy brief. Working paper
- Pandit A, Jeong H, Crittenden JC, Xu M, Ariaratnam S, Begovic M, Bras B, Brown M, DesRoches R, French S, Georgakakos A, Karady G, Li K, Noonan D, Perrings C, Williams E (2010) Sustainable and resilient urban infrastructure design: a holistic approach. In: The 6th international conference on sustainable water environment, session 4.7: Infrastructure, Newark, DE
- Gagnon L (2002) Life-cycle assessment of electricity generation options: the status of research in 2001. *Energy Policy* 30(14):1267–1278
- Georgia Power (n.d.) Plant Branch. Retrieved from Georgia Power. www.georgiapower.com/about/pdf/Plant%20branch%20Brochure.pdf
- Georgia Power (n.d.) Plant Branch. Retrieved from Georgia Power. www.georgiapower.com/about/pdf/Plant%20branch%20Brochure.pdf
- Georgia Power (n.d.) Plant Robert W. Scherer. Retrieved 2010, from Georgia Power. www.georgiapower.com/about/pdf/Plant%20sherer%20Brochure.pdf
- Georgia Power (n.d.) Plant Wansley. Retrieved from Georgia Power. www.georgiapower.com/about/pdf/Plant%20Wansley%20Brochure.pdf
- US Energy Information Administration (2010) Updated capital cost estimates for electricity generation plants
- Miroslav M, Begovic IK (2011) Distributed renewable PV generation in urban distribution networks. In: PSCE 2011, IEEE power systems conference and exposition, Phoenix

36. Borin S, Wang J, Thomas V (2009) City of Atlanta greenhouse gas emissions inventory. Report to the City of Atlanta, Atlanta
37. Westerhoff P, Crittenden J (2009) Urban infrastructure and use of mass balance models for water and salt. In: LA Baker, ed. *The Water Environment of Cities*. Boston, MA: Springer US, pp 49–68
38. Brown MA, Southworth F, Sarzynski A (2008) Shrinking the carbon footprint of metropolitan America. The Brookings Institution, Washington, DC
39. Scott C, Pasqualetti M, Hoover J, Garfin G, Varady R, Guhathakurta S (2009) Water and energy sustainability with rapid energy growth and climate change in the Arizona-Sonora border region. Report to the Arizona water institute (closed)

Books and Reviews

- Agence France-Presse (2006) Heat wave claims lives in Europe, and hottest days are still to come. *New York Times-Europe*, 19 July 2006
- US EPA Region 2 (2006) Report on the City of New York's progress in implementing the watershed protection program, and complying with the filtration avoidance determination. US EPA Region 2. New York City watershed team and New York State Department of Health Bureau of water supply protection. US EPA, Washington, DC
- Tugend A (2008) If your appliances are avocado, they probably aren't green. *New York Times*, 12 May 2008
- Klein G (2009) The water-energy-carbon footprint connection
- Wei M, Patadia S, Kammen D (2010) Putting renewables and energy efficiency to work: how many jobs can clean energy industry generate in the US? *Energy Policy* 38(2):919–931
- Bruno L (1931) Magic plates, tap sun for power. *Pop Sci* 118:41
- Torcellini P, Long N, Judkoff R (2003) Consumptive water use for U.S. power production. NREL, Golden, p 5, Dec 2003
- Annual Energy Review (2009) U.S. Energy Information Administration. US DOE, Washington, DC, p 230, Aug 2010
- Lazard Ltd (2009) Levelized cost of energy analysis, version 3.0. Lazard, New Orleans, p 2, Feb 2009

Energy Storage: Ultracapacitor

ANDREW F. BURKE

Institute of Transportation Studies, University of California-Davis, Davis, CA, USA

Article Outline

Glossary

Definition

Introduction

Energy Storage Characteristics of Ultracapacitors and Batteries

Systems Considerations and Advantages of Combining Ultracapacitors and Batteries

Laboratory and Vehicle Field Experiences Using Ultracapacitors

Hybrid Vehicle Simulations Using Ultracapacitors Alone

Plug-in Hybrid Vehicle Simulations Using Advanced Batteries And Ultracapacitors

Fuel Cell Vehicle Simulations Using Ultracapacitors Economic Considerations

Future Directions of Ultracapacitor Development and Vehicle Applications

Summary

Bibliography

Glossary

Battery A cell in which electrical energy is stored and discharged via Faradaic electrochemical reactions occurring at a predetermined voltage dependent on the active materials in the two electrodes.

Control (power split) strategy The strategy used in a hybrid-electric vehicle to determine the fraction of the power demanded that is provided by the electric motor and engine, respectively.

Hybrid-electric vehicle (HEV) A vehicle having both an electric motor and internal combustion engine in which the battery is maintained near a fixed state-of-charge via electrical energy generated by the electric motor/generator driven by the engine.

Plug-in hybrid electric vehicle (PHEV) A vehicle having both an electric motor and internal combustion engine in which the battery is depleted by all-electric driving and recharged from the wall plug.

Pulse power efficiency The fraction of the discharge or charge energy from the battery or ultracapacitor during a short, high-power pulse that is available as electrical energy for use by the vehicle.

Ultracapacitor (electrochemical capacitor) A device in which electrical energy is stored and discharged via the electrostatic separation of charge in the micropores of activated carbon in two electrodes.

Vehicle simulations Computer calculations of the operation of a vehicle based on models of the various components in the driveline of the vehicle and a specified control strategy for its operation.

Definition

Most hybrid vehicles use batteries for energy storage. The batteries must provide both the energy and power required by the vehicle. In all cases, this requires compromises in designing the battery in that it cannot be optimized for either energy density or power capability. In addition, in most cases the battery unit is oversized (stores more energy than is used by the vehicle) in order to achieve long cycle life even for the shallow cycles experienced in charge-sustaining hybrid vehicles like the Toyota Prius. As discussed in this entry, another approach to providing energy storage for hybrid vehicles, both charge-sustaining and plug-in designs, is to utilize ultracapacitors either alone or in combination with batteries. In this approach, the batteries can be optimized for energy density and cycle life and ultracapacitors can provide the power both for acceleration and regenerative braking. The ultracapacitors, which are inherently high-power devices, would be deep discharged to one-half rated cell voltage and still provide cycle life of 500,000 to one million cycles.

The development of ultracapacitors suitable for use in vehicle powertrains started in about 1990 and steady progress in improving the performance of the devices has occurred since that time. This entry is concerned with the present status of commercially available cells and modules and projections of future developments based on the performance of prototype cells of both the carbon/carbon and hybrid carbon/pseudo-capacitive designs. The cost of ultracapacitors is currently the primary obstacle to the use of ultracapacitors in vehicle applications. In this entry, the future costs are projected and compared with high-power batteries based on the ultracapacitor material costs. Finally, applications of ultracapacitors in hybrid-electric vehicle powertrains are discussed. Key considerations are the relationships between the Wh stored in the ultracapacitor unit, system control strategy, and the fuel economy improvements attained for various driving cycles. These relationships are studied using detailed vehicle simulation results from Advisor.

Introduction

In the first section of the entry, the energy storage characteristics of lithium batteries and ultracapacitors are presented and compared. Of particular interest is

the comparative power capability of lithium batteries and carbon/carbon ultracapacitors for charge/discharge conditions to be encountered in hybrid-electric vehicles. In the next section, system considerations and advantages of combining ultracapacitors and batteries for energy storage are discussed. Next laboratory and field experiences with ultracapacitors in hybrid-electric vehicles are reviewed for transit bus, light-duty, and fuel cell applications.

The final sections of the entry discuss in detail simulations of vehicles on various driving cycles using ultracapacitors alone and in combination with advanced batteries. The vehicle simulations treat charge-sustaining micro- and mild-hybrid designs, plug-in hybrids, and fuel cell-powered vehicles. The charge-sustaining hybrid vehicles using ultracapacitors alone in place of batteries are discussed in terms of the sizing of the ultracapacitor unit and control strategies that optimize the fuel economy improvements achievable. For the plug-in hybrids, the emphasis is on the use of ultracapacitors with advanced batteries having energy densities greater than 200 Wh/kg. It is likely that these batteries will not have proportionally high power density and will require the assist of ultracapacitors to achieve all-electric operation even on the FUDS driving cycle. The next section of the entry deals with fuel cell-powered vehicles and considers the use of ultracapacitors to load level the fuel cell to increase system efficiency and thus improve hydrogen consumption. In the final section of the entry, the cost of batteries and ultracapacitors are compared for charge-sustaining hybrid applications.

Energy Storage Characteristics of Ultracapacitors and Batteries

Lithium batteries and ultracapacitors have been tested in the laboratory at the University of California, Davis [1–3]. A summary of the test results for the batteries is given in Table 1 and for ultracapacitors in Table 2. For both energy storage technologies, the devices with the highest energy density typically have the lowest power capability. The pulse power capabilities shown in the tables were calculated using the following relationships:

$$\text{Batteries: } P = EF(1 - EF) V_{oc}^2/R$$

$$\text{Ultracapacitors: } P = 9/16(1 - EF) V_{rated}^2/R$$

Energy Storage: Ultracapacitor. Table 1 Summary of the performance characteristics of lithium-ion batteries of various chemistries

Battery developer/ cell type	Electrode chemistry	Voltage range	Ah	Resist. (mOhm)	Wh/kg	W/kg 90% eff. ^a	W/kg match. imped.	Wt. (kg)	Density (gm/cm ³)
Enerdel HEV	Graphite/ Ni MnO ₂	4.1–2.5	15	1.4	115	2,010	6,420	0.445	–
Enerdel EV/PHEV	Graphite/ Ni MnO ₂	4.1–2.5	15	2.7	127	1,076	3,494	0.424	–
Kokam prismatic	Graphite/ NiCoMnO ₂	4.1–3.2	30	1.5	140	1,220	3,388	0.787	2.4
Saft cylind.	Graphite/ NiCoAl	4.0–2.5	6.5	3.2	63	1,225	3,571	0.35	2.1
GAIA cylind.	Graphite/ NiCoMnO ₂	4.1–2.5	40	0.48	96	2,063	5,446	1.53	3.22
			7	3.6	78	1,310	3,472	0.32	–
A123 cylind.	Graphite/ iron phosph.	3.6–2.0	2.2	12	90	1,393	3,857	0.07	2.2
Altairnano prismatic	LiTiO/ NiMnO ₂	2.8–1.5	11	2.2	70	990	2,620	0.34	1.83
Altairnano prismatic	LiTiO/ NiMnO ₂	2.8–1.5	3.8	1.15	35	2,460	6,555	0.26	1.91
Quallion cylind.	Graphite/ NiCo	4.2–2.7	1.8	60	144	577	1,550	0.043	2.6
Quallion cylind.	Graphite/ NiCo	4.2–2.7	2.3	72	170	445	1,182	0.047	2.8
EIG prismatic	Graphite/ NiCoMnO ₂	4.2–3.0	20	3.1	165	1,278	3,147	0.41	–
EIG prismatic	Graphite/ iron phosph.	3.65–2.0	15	2.5	113	1,100	3,085	0.42	–
Panasonic EV prismatic	Ni metal hydride	7.2–5.4	6.5	11.4	46	395	1,093	1.04	1.8

^apower density $P = \text{Eff.} \cdot (1 - \text{Eff.}) \cdot V_{oc}^2 / R$, $P_{\text{match. imped.}} = V^2 / 4R$

where EF is the efficiency of the pulse ($EF = V_{\text{pulse}} / V_{oc}$). The matched impedance power which is often cited for both battery and ultracapacitor devices is calculated as follows:

$$P_{\text{matchimped.}} = V^2 / 4R$$

These power relationships are derived in [4].

For charge-sustaining hybrids, it seems reasonable to cite the power capability of devices for pulse power efficiencies of 90–95%. For PHEVs and EVs operating

in charge-depleting modes for the battery, it is reasonable to cite the power for efficiencies of 75–80%. In all instances, the power capability is proportional to V^2/R indicating that high power capability requires a low resistance R .

Comparisons of the Power Capabilities of Lithium Batteries and Ultracapacitors

There has been considerable discussion in the literature [4–6] comparing the power capability of lithium-ion

Energy Storage: Ultracapacitor. Table 2 Summary of ultracapacitor device characteristics

Device	V rated	C (F)	R (mOhm)	RC (s)	Wh/kg (1)	W/kg (95%) (2)	W/kg match. imped.	Wt. (kg)	Vol. (l)
Maxwell ^a	2.7	2,885	0.375	1.08	4.2	994	8,836	0.55	0.414
Maxwell	2.7	605	0.90	0.55	2.35	1,139	9,597	0.20	0.211
Skeleton technologies	2.8	1,600	1.3	2.1	5.8	800	7,140	0.22	0.13
Yunasko ^b	2.7	55	4	0.22	5.5	5,695	50,625	0.009	–
Yunasko ^b	2.7	450	1.3	0.58	5.89	2,766	24,595	0.057	0.045
Yunasko ^b	2.7	510	0.9	0.46	5.0	2,919	25,962	0.078	0.055
Ness	2.7	1,800	0.55	1.00	3.6	975	8,674	0.38	0.277
Ness	2.7	3,640	0.30	1.10	4.2	928	8,010	0.65	0.514
Ness (cyl.)	2.7	3,160	0.4	1.26	4.4	982	8,728	0.522	0.379
Asahi glass (propylene carbonate)	2.7	1,375	2.5	3.4	4.9	390	3,471	0.210 (estimated)	0.151
Panasonic (propylene carbonate)	2.5	1,200	1.0	1.2	2.3	514	4,596	0.34	0.245
EPCOS	2.7	3,400	0.45	1.5	4.3	760	6,750	0.60	0.48
LS Cable	2.8	3,200	0.25	0.80	3.7	1,400	12,400	0.63	0.47
BatScap	2.7	2,680	0.20	0.54	4.2	2,050	18,225	0.50	0.572
Power sys. (activated carbon, propylene carbonate) ^b	2.7	1,350	1.5	2.0	4.9	650	5,785	0.21	0.151
Power sys. (graphitic carbon, propylene carbonate) ^b	3.3	1,800	3.0	5.4	8.0	486	4,320	0.21	0.15
	3.3	1,500	1.7	2.5	6.0	776	6,903	0.23	0.15
JSR micro (AC/graphitic carbon)	3.8	1,000	4	4	11.2	900	7,987	0.113	0.073
		2,000	1.9	3.8	12.1	1,038	9,223	0.206	0.132

(1) Energy density at 400 W/kg constant power, $V_{rated} - 1/2 V_{rated}$

(2) Power based on $P = 9/16 * (1 - EF) * V^2 / R$, EF = efficiency of discharge

^aExcept where noted, all the devices use acetonitrile as the electrolyte

^bExcept for these, all devices are packaged in metal containers

batteries and ultracapacitors. The conclusions vary from statements that lithium batteries have power capability equal to that of ultracapacitors to statements that ultracapacitors have an order of magnitude higher power capability than lithium batteries. Detailed comparisons of the power capability of ultracapacitors and batteries taken from [4] are shown in Table 3. As indicated in the table, neither of the extreme statements is valid in general and that comparisons should be made between specific devices for specific applications.

Comparisons are often made based on the matched impedance power of the two types of devices. These comparisons indicate that most ultracapacitors have a power capability (W/kg) of three to six times that of lithium batteries. However, for vehicle applications the matched impedance power is not appropriate and should not be the basis of comparison.

For charge-sustaining HEV applications, a good basis of comparison is the W/kg at 95% efficiency at the SOC at which the devices will be used in the

Energy Storage: Ultracapacitor. Table 3 Comparisons of the power capabilities of various devices for HEVs and PHEVs using the different methods of calculation

Device	Matched impedance	USABC Min/max	Efficient pulse EF =95%	Efficient pulse EF =80%
Lithium batteries 60% SOC				
Kokam NCM 30 Ah	2,893	2,502	550	1,848
Enerdel HEV	5,491	4,750	1,044	3,507
NCM 15 Ah				
Enerdel EV	2,988	2,584	568	1,908
NCM 15 Ah				
EIG NCM	2,688	2,325	511	1,721
20 Ah				
EIG FePhosph.	2,141	2,035	458	1,540
15 Ah				
Altairnano LiTiO	1,841	1,750	350	1,180
11 Ah				
Altairnano LiTiO	4,613	4,385	992	3,341
3.8 Ah				
Ultracapacitors				
$V_0 = 3/4V_{\text{Rated}}$				
Maxwell 2,890 F	8,836	4,413	994	
Nesscap 3,100 F	8,730	4,360	982	
BatScap 2,700 F	18,224	9,102	2,050	
Yunasko 450 F	22,838	11,406	2,569	
LS Cable 3,200 F	12,446	4,609	1,038	
JSR 2,000 F	9,228	6,216	1,400	

vehicle. On this basis, there are lithium batteries with the same power capability as some carbon/carbon ultracapacitors, but there are some ultracapacitors with power capability twice that of the highest power lithium batteries presently available for vehicle applications. In other words, it is not possible to make general statements that are applicable to all devices of either type. The issue is further complicated when one notes that the density of the lithium batteries is about twice that of carbon/carbon ultracapacitors (2.2 g/cm^3 for the batteries and 1.2 g/cm^3 for the ultracapacitors). Hence on a volume basis W/l at 95% efficiency, the differences between the batteries and ultracapacitors

are often quite small. Hence for HEVs, batteries alone and ultracapacitors alone can be an option with the decision being based on cycle life and cost in addition to relative power capability [7].

For plug-in hybrid and battery electric vehicle applications, the maximum usable power density from the lithium-ion battery can be higher than in an HEV because the peak power of the driveline is used less frequently and consequently charge/discharge efficiency is less important. For example, a pulse power efficiency of 80% is probably sufficient and most of the lithium batteries have a power capability of greater than 1,000 W/kg, 2,200 W/l for that efficiency.

In addition, the battery is larger (heavier) in these vehicles and as a result, the power density requirement is less demanding. For PHEVs and EVs, the best application of ultracapacitors is likely to be in combination with batteries designed for high energy density, long cycle life, and low cost. In those cases, as discussed in a later section of the entry, the ultracapacitors greatly reduce the peak currents and dynamic stress on the batteries and thus extend their cycle life.

Systems Considerations and Advantages of Combining Ultracapacitors and Batteries

Vehicle Design Considerations

The energy storage requirements vary a great deal depending on the type and size of the vehicle being designed and the characteristics of the electric powertrain to be used. Energy storage requirements for various vehicle designs and operating modes are shown in Table 4 for a midsize passenger car. Requirements are given for electric vehicles and both charge-sustaining and plug-in hybrids. These requirements can be utilized to size the energy storage unit in the vehicles when the characteristics of the energy storage cells are known. In some of the vehicle designs considered in Table 4, ultracapacitors are used to provide the peak power rather than batteries.

In the cases using only ultracapacitors, the key issue is the minimum energy (Wh) required to operate the

vehicle in real-world driving because the energy density characteristics of ultracapacitors are such that the power and cycle life requirements will be met in most cases if the unit is large enough to meet the energy storage requirement. As shown in Table 4, for passenger car applications, the energy storage in the ultracapacitor can be 150 Wh or less even if the ultracapacitor is used alone for the energy storage.

Reasons for Combining Ultracapacitors and Batteries

Most hybrid vehicles use batteries for energy storage. The batteries must provide both the energy and power required by the vehicle. In all cases, this requires compromises in designing the battery in that it cannot be optimized for either energy density or power capability. In addition, in most cases the battery is oversized (stores more energy than is used by the vehicle) in order to achieve long cycle life even for the shallow cycles experienced in charge-sustaining hybrid vehicles like the Toyota Prius. Combining ultracapacitors with batteries permits the design of the batteries to be optimized for energy density and cycle life and limits the magnitude and frequency of maximum power pulses experienced by the batteries during vehicle operation. The result will be a lighter, smaller, and lower cost battery which will have much longer cycle life (by as much as a factor of 2–3) and will operate over

Energy Storage: Ultracapacitor. Table 4 Energy storage unit requirements for various types of electric drive midsize passenger cars

Type of electric driveline	System voltage V	Usable energy storage	Maximum pulse power at 90–95% efficiency (kW)	Cycle life (number of cycles)	Usable depth-of-discharge
Electric	300–400	15–30 kWh	70–150	2,000–3,000	deep
					70–80%
Plug-in hybrid	300–400	6–12 kWh battery	50–70	2,500–3,500	deep
		100–150 Wh			60–80%
		ultracapacitors			
Charge-sustaining hybrid	150–200	100–150 Wh	25–35	300 K–500 K	Shallow
		ultracapacitors			5–10%
Micro-hybrid	45	30–50 Wh	5–10	300 K–500 K	Shallow
		ultracapacitors			5–10%

significantly narrower ranges of voltage and temperature than would be the case using only a battery. In addition, the cooling requirements and complexity of the battery management unit of the battery will be reduced and be less costly.

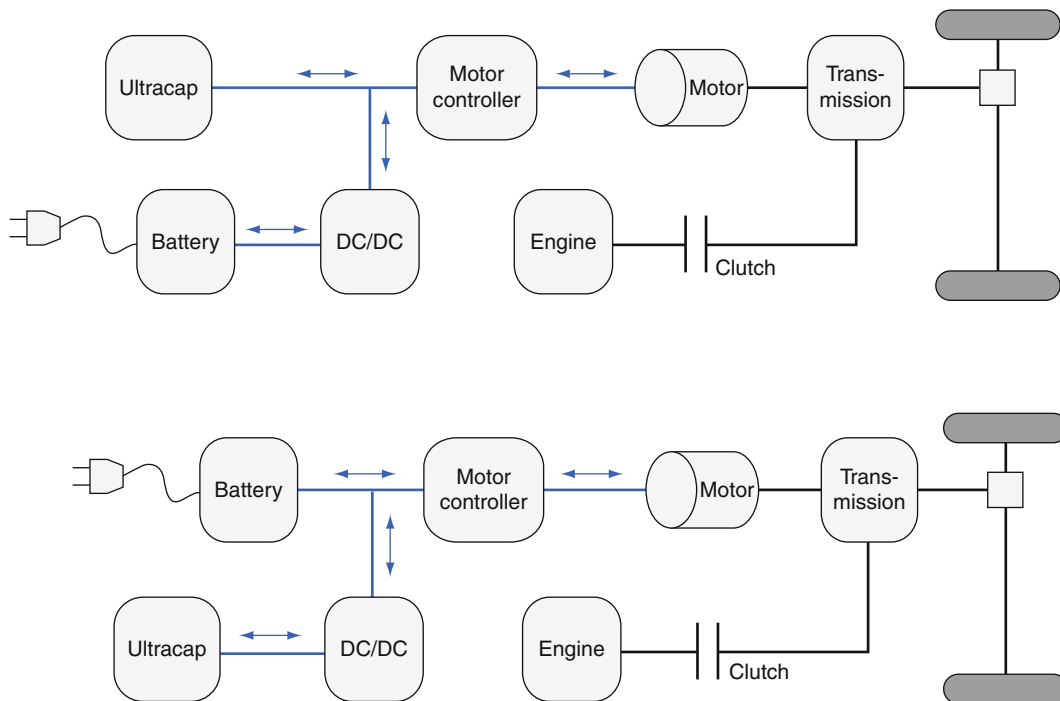
The ultracapacitors would provide the high-power pulses in discharge and accept all the charge pulses during regenerative braking of the vehicle. If needed, the ultracapacitors can be recharged at high rates during regenerative braking and/or using power from the engine or fuel cell. The ultracapacitors would be deep discharged (up to 75% of their rated energy) when that energy is needed to power the vehicle. Even for deep discharge operation, the ultracapacitors can be expected to have a cycle life of at least 500,000 cycles. As noted in Table 4, the energy stored (Wh) in the ultracapacitors can be quite limited being less than 150 Wh for midsize passenger cars. Larger ultracapacitor units would be needed for buses and trucks, but their advantages would be similar to those used in passenger cars. The weight and cost of the ultracapacitor unit would be proportional to the

energy stored so it is advantageous to limit the energy stored to the minimum required for a particular application. As shown in Fig. 1, for most applications, it will be necessary to utilize interface electronics to control the power flow from the ultracapacitors because their voltage changes significantly during discharge. This permits a large fraction of the usable energy of the ultracapacitors to be utilized and minimizes their cost.

Sizing Components: Weight and Volume

Whether or not a vehicle designer will consider using ultracapacitors alone or in combination with batteries instead of a battery pack for energy storage will depend to a large extent on the relative weight, volume, and cost of the two systems.

Consider first the case of charge-sustaining hybrid vehicles (HEVs). In this case, the energy storage unit is sized by both usable power (kW) and energy storage (Wh) requirements. For batteries, the key issues are the power requirement and the minimum usable energy consistent with high cycle life for shallow cycles.



Energy Storage: Ultracapacitor. Figure 1

Schematics of powertrains using ultracapacitors and batteries

The total energy stored in the battery unit is of secondary importance as far as operation of the vehicle is concerned, but it has a strong effect on the weight, volume, and cost of the unit. For ultracapacitors, the key issue is the minimum energy (Wh) required to operate the vehicle in real-world driving because the energy density characteristics of ultracapacitors are such that the power and cycle life requirements will be met if the unit is large enough to meet the energy requirement.

Energy storage units for charge-sustaining hybrid powertrains for midsize passenger cars are shown in Table 5. For each application, the usable power and energy, total energy stored, and incremental

state-of-charge for the unit and its weight and volume are shown in the table. The results shown in Table 5 indicate that for charge-sustaining hybrids, either batteries or ultracapacitors can be used alone. All the units listed should function satisfactorily for the applications indicated with the decision concerning their relative attractiveness based to a large extent on cost, which at the present time is uncertain for all the systems. The results in Table 5 also indicate that a combination of lead-acid batteries and ultracapacitors would result in a relatively low weight unit for a micro-hybrid having a maximum electric motor power of 5–10 kW. The economics of such a combination should be attractive because of the low cost of the lead-acid batteries.

Energy Storage: Ultracapacitor. Table 5 Characteristics of various technologies/types of batteries for use in charge-sustaining hybrid vehicle applications

<i>Charge-sustaining hybrids (HEVs)</i>								
Technology	C or Ah	System voltage	Δ SOC%	Total energy Wh ^a	Max. Power kW ^b	Pulse efficiency%	Weight of cells kg	Volume of cells l
Ultracapacitors								
Carbon/carbon	2,400	160	75	100	>40	90	22	15
microporous C/graphitic C	2,400	160	75	150	27	90	18	13
Lithium-ion batteries								
Graphite/Ni Co Al	9	160	6.6	1,500	31	91	24	11
Iron phosphate 2.2 Ah cells	11	160	5.5	2,000	27	91	23	11
Lithium titanate 11 Ah	11	160	5.5	1,800	27	90	26	14
Lithium titanate 2.5 Ah	4	160	16.6	600	27	90	16	9
Nickel metal hydride (Prius)					14	90		
	10	160	6.25	1,600	25	80	35	19
Micro-hybrids								
Lead-acid	50	45	5	1,800	5	90	53	19
Lead-acid/ultracapacitors	28		10	1,000			29	10
	4,000	45	75	50	15	95	10	8
							39 total	18 total
Nickel metal hydride (Prius)	24	45	10	1,100	10	90	24	13

^aTotal energy stored

^bPower at the pulse efficiency indicated

Next consider energy storage units for plug-in hybrid vehicles (PHEVs). A key design parameter for PHEVs is the all-electric range. Energy storage units will be considered for all-electric ranges of 10, 15, 20, 30, and 40 miles. The acceleration performance of all the vehicles will be the same (0–60 mph in 8–9 s). For the batteries, the usable depth of discharge will be taken to be 70%. For the ultracapacitors, it is assumed that they can be cycled from rated to one-half rated voltage using 75% of the total stored energy. In the PHEVs, the battery will be sized by the energy needed to sustain a specified all-electric range. Hence the weight and volume of the battery pack follow directly from its energy density (Wh/kg, Wh/l). This means that battery technologies with high energy density will be strongly favored for PHEV applications. However, the batteries must also be able to meet the power requirements of the PHEV powertrain which can approach those of the battery-powered EV. Unfortunately, battery designs to attain the maximum energy density in most cases require a sacrifice in power capability as shown in Table 1. Calculated power density requirements for PHEVs of various ranges and maximum power [8] are given in Table 6 for batteries having energy densities of 70, 100, and 200 Wh/kg. The values shown are for a midsize passenger car having an energy usage of 250 Wh/mile in the all-electric mode and maximum electric motor power of 50 and 70 kW.

The required power densities for all-electric ranges of 20 miles and less are high especially for high energy density batteries. When the power density required

exceeds that of the battery, it is of interest to consider using an ultracapacitor to augment the power of the battery. The usable pulse power from a battery depends on the efficiency that is acceptable for a particular application. The power densities for efficiencies of 95% and 80% for the various battery technologies are shown in Table 3. If it is assumed that 80% efficiency is satisfactory for use of the battery in the charge-depleting mode of the PHEV, it can be concluded from Tables 3 that ultracapacitors would be needed by the 200 Wh/kg lithium-ion and the 70 Wh/kg nickel metal hydride batteries. This result is shown in Table 7. The energy storage unit weights for the PHEVs of various all-electric ranges and battery energy density are given in Table 8. The combination of ultracapacitors and 200 Wh/kg lithium-ion batteries have the lowest weight for all the ranges. For higher energy density batteries (>250 Wh/kg), it seems even more likely that ultracapacitors will be needed to meet the power demands of PHEVs. The combination of ultracapacitors and batteries will also yield the highest efficiency during the charge-sustaining mode of vehicle operation after the energy of the batteries has been depleted.

As noted above, ultracapacitors can be used with 70 Wh/kg nickel metal hydride batteries for PHEV applications due to their limited power capability. Even using ultracapacitors, the all-electric range with nickel metal hydride batteries is limited to about 15 miles as otherwise the weight of the energy storage unit would exceed 100 kg.

Energy Storage: Ultracapacitor. Table 6 Battery sizing and power density for various ranges and motor power

Range miles	kWh ^a needed	kWh ^b stored	Battery 200 Wh/kg			Battery 100 Wh/kg			Battery 70 Wh/kg		
			^b kg	50 kW kW/kg	70 kW kW/kg	kg	50 kW kW/kg	70 kW kW/kg	kg	50 kW kW/kg	70 kW kW/kg
10	2.52	3.6	18	2.78	3.89	36	1.39	1.94	51	0.98	1.37
15	3.78	5.4	27	1.85	2.59	54	0.92	1.30	77	0.65	0.91
20	5.04	7.2	36	1.39	1.94	72	0.69	0.97	103	0.49	0.68
30	7.56	10.8	54	0.93	1.30	108	0.46	0.65	154	0.32	0.46
40	10.1	14.4	72	0.69	0.97	144	0.35	0.49	206	0.24	0.34

^aVehicle energy usage from the battery: 250 Wh/mi

^bUsable state-of-charge for batteries: 70%, weights shown are for cells only

Energy Storage: Ultracapacitor. Table 7 Battery/ultracapacitor combinations for various all-electric ranges

Wh/kg	100	200	100	90	70	70
Range miles	GAIA	Quallion 200	Quallion 100	A123	Altairnano	Panasonic Nickel mt.hyd.
10	N/N ^a	Y/Y	N/N	N/N	N/N	Y/Y
15	N/N	Y/Y	N/N	N/N	N/N	Y/Y
20	N/N	Y/Y	N/N	N/N	N/N	Y/Y
30	N/N	Y/Y	N/N	N/N	N/N	Y/Y
40	N/N	N/N	N/N	N/N	N/N	Y/Y

^aX/X means need at 50 kW/need at 70 kW yes Y or no N

Energy Storage: Ultracapacitor. Table 8 Storage unit weights using a combination of batteries and ultracapacitors for various all-electric ranges

Wh/kg	5	200		100		70	
Range miles	Ultracap kg ^a	Battery Kg ^b	Combination kg	Battery kg	Combination kg	Battery kg	Combination kg
10	20	18	38	36	56	51	71
15	20	27	47	54	74	77	97
20	20	36	56	72	92	103	123
30	20	54	74	108	128	154	174
40	20	72	92	144	164	206	226

^aThe ultracapacitor unit stores 100 Wh usable energy

^bWeights shown are for cells only, packaging into modules not Included

Power Split Strategies

As noted previously, interface electronics is needed (Fig. 1) in hybrid powertrains using a combination of batteries and ultracapacitors to control the power split between the two energy storage units. Selection of a control strategy that will both limit the maximum power demanded of the batteries and improve the efficiency of the powertrain is critical to the design of attractive HEVs and PHEVs. One approach for the PHEV applications is to have the battery meet the average vehicle power requirement and the ultracapacitor meet the variable, pulse power requirement. The average vehicle power demand can be calculated over some specified time period (30–120 s) depending on the driving conditions. As discussed later, vehicle simulations indicate that this approach results in relatively slow

variations in battery power and should lead to significantly increased battery cycle life and reduced need for high-power cells. Using this approach, the interface electronics could be directly attached to either the batteries or ultracapacitors (see Fig. 1).

As in the case of the HEVs, the key element of the system control strategy is how the ultracapacitors are recharged utilizing engine power via the motor/generator. The vehicle simulations have shown that the ultracaps should be recharged during periods in which the engine is needed to meet the vehicle power requirements. In this way, the engine operates nearly all the time at high efficiency and not at low output torque conditions resulting in large improvements (>35%) in fuel economy compared to the reference engine alone case.

Laboratory and Vehicle Field Experiences Using Ultracapacitors

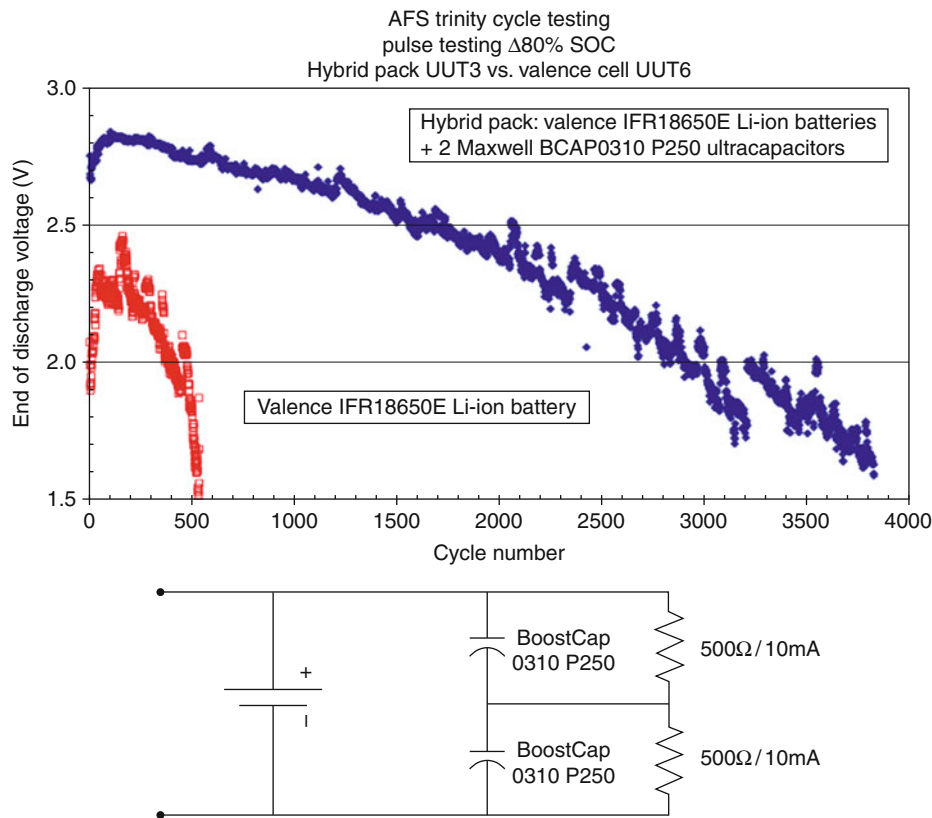
Laboratory Testing of Ultracap/Battery Combinations

Most of the research [9–11] involving the testing of ultracap/battery combinations has been performed on bench-top arrangements. In most cases, these tests have included state-of-the-art interface electronics that controlled the currents from the ultracapacitors. The intent of these studies has been to show that the stress on the batteries due to repeated high current pulses was greatly reduced using ultracapacitors resulting in extended battery cycle life. The test data show clearly that the mean currents from the batteries are greatly reduced, but the effect on battery cycle life is not yet clear from the data. Much further testing of this type, especially using lithium batteries, is needed. The test results [12] from a simpler test utilizing small

spiral wound lithium batteries (1.4 Ah) and 310 F ultracapacitors without interface electronics are shown in Fig. 2. The data show the effect of load leveling on the cycle life of the cells which were pulsed at the 5C rate. The magnitude of the effect (a factor of 7–8) is likely much greater than can be expected with larger batteries, but nevertheless, the data do indicate that load-leveling batteries can be expected to significantly improve cycle life.

Vehicle Field Experience Using Ultracapacitors

There has been continuing experience in demonstrating the use of ultracapacitors in hybrid-electric powertrains in heavy-duty [13–15] and light-duty [16–18] vehicles. In these applications, the capacitors provide the peak power for acceleration and recover energy during periods of braking. An engine generator or fuel cell is used to generate onboard electricity to



Energy Storage: Ultracapacitor. Figure 2

Pulsed discharge at average rate of 1 C comprising 5 C 5 s pulses at a 20% duty factor. Each cycle 80% (1.12 Ah)

recharge the capacitors as needed. The control strategy for the hybrid vehicle should be designed to operate the engine or fuel cell as close to the maximum efficiency condition as the driving cycle permits. Capacitor units for use in hybrid vehicles must be high voltage as the peak power for these applications is high (up to 200 kW). For passenger cars and SUVs, the voltage is likely to be about 300 V and for transit buses about 700 V. This would require 125 cells in series for the light-duty vehicles and 250 cells in series for the buses.

Transit Bus Applications In the USA, the company most active in utilizing ultracapacitors in hybrid-electric powertrains for buses and large trucks has been the ISE Corporation in San Diego, California [13]. ISE has developed a 360 V capacitor unit consisting of 144 2,600 F Maxwell cells connected in series (see Fig. 3). The weight and volume are 114 kg and 189 l, respectively. The unit stores 0.325 kWh of energy (0.245 kWh usable). In a transit bus, two of the units are used in series resulting in a voltage of 720 V and energy storage of 0.650 kWh. The peak power

capability of the combined unit is over 300 kW. ISE utilizes this capacitor unit with a 225 kW electric motor in series hybrids using gasoline and diesel engines and hydrogen fuel cells. Since the capacitor unit stores only about 0.5 kWh, it can provide power only during vehicle acceleration and recovers energy during braking. The engine or fuel cell must provide all the power during cruise and hill climbing. ISE has built over 100 buses using the ultracapacitor energy storage units for transit companies in Southern California. The buses are in daily revenue service. Field experience has shown that the capacitor unit functions as planned with a high reliability. Field operation of the buses has shown a fuel economy improvement of 25% compared to the standard diesel engine buses [13].

Hybrid-Electric Passenger Car Applications Unlike the transit bus applications described in the previous section, there has not been a commercialized use of ultracapacitors in passenger cars as of 2010. In the last couple of years, there has been considerable discussion [19, 20] of marketing of micro-hybrids with stop-start



Nominal voltage	360V*
Peak voltage	403V**
Rated current	400A
Capacitance	18.05F
Total energy stored nominal/max.	0.325k Wh/0.407k Wh
Leakage current	5mA nominal
Operating temperature	−35 to 65°C
Weight	240lbs
Dimensions w×l×h	24"×40"×12"
Standard pack	144 capacitors
Fire suppression system	Heat activated halotron system

Energy Storage: Ultracapacitor. Figure 3
The ISE ultracapacitor unit (360 V, 0.325 kWh)

capability to improve fuel economy particularly in Europe. It is likely that when these cars are finally marketed some of them will use ultracapacitors. There have been several noteworthy demonstrations of ultracapacitors in hybrid-electric passenger cars. As discussed in the following paragraphs, these projects were very successful and showed that ultracapacitors can be used in place of batteries in mild hybrid vehicles with fuel economy improvements at least as high as with batteries.

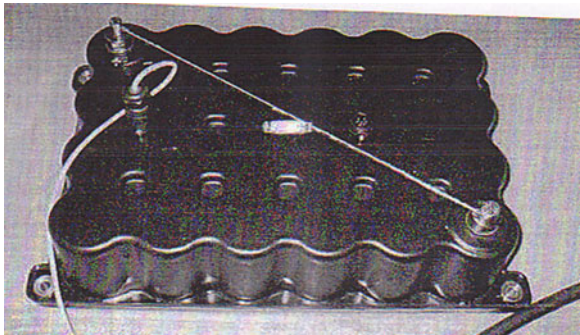
The first project was termed “SUPERCAR” and was funded by the European Community (EC). It was a joint project between EPCOS [16], the ultracapacitor developer, and Siemens VDO [21], the vehicle integrator. The parallel hybrid passenger car (VW Golf) combined the ultracap module and lead-acid battery into a 42 V, 10 kW (peak) electric driveline with a 66 kW engine. The vehicle was tested on both the chassis dynamometer and the road. The tests showed a 16–18% improvement in fuel economy compared to the standard ICE car. The capacitor unit consisted of 24 3,600 F cells connected in series (see Table 2 for the cell characteristics). The rated voltage of the unit is 60 V and its capacitance and resistance are 150 F and 8 mOhm, respectively. The unit is shown in Fig. 4.

A second, more recent project (2009), is described in detail in [17]. This project was performed by the National Renewable Energy Laboratory (NREL) and General Motors (GM). It involved installing ultracapacitors in a 2007 Saturn Vue hybrid along with the nickel metal hydride battery. The vehicle could be operated on either of the energy storage units alone in order to make a direct comparison of

its operation on the ultracapacitors and battery. The ultracapacitor energy storage unit consisted of one or two 48 V, 165 F modules from Maxwell. Each module, which consisted of 18 3,000 F cells connected in series (see Table 2 for the characteristics of the cells), stored about 35 Wh. A special UCAP state estimator was utilized to maintain the ultracapacitors in the required range of state-of-charge.

The 2007 Saturn Vue hybrid is a mild hybrid using a belt alternator/starter in parallel with a 127 kW, variable valve timing gasoline engine. The alternator/starter electric machine has a continuous rating of 3 kW and a peak power of 6–8 kW. The nickel metal hydride battery has a nominal rating of 36 V, but vehicle testing indicated it operates in the 38–42 V range most of the time at 75% state-of-charge. The characteristics of the energy storage units are compared in Table 9.

Extensive dynamometer testing of the vehicle was done operating on the ultracapacitors (both with one module and two modules) and the nickel metal hydride battery. The vehicle was tested on several driving cycles (FUDS, FHWD, and US06). The three energy storage options could be utilized interchangeably using the same vehicle operating strategy. The vehicle operated essentially the same on each of the energy storage options. The testing indicated that only 25–30 Wh was used from the ultracapacitor units even on the US06 driving cycle. When using only one ultracap module on the FUDS cycle, the state-of-charge (V/V_{rated}) of the ultracap unit varied from 75–90%. There was no difference in the measured fuel economy when the vehicle was operated on the ultracapacitors or the nickel metal hydride battery for any of the three



Energy Storage: Ultracapacitor. Figure 4
The EPCOS 60 V ECC module (150 F, 56 Wh)

Energy Storage: Ultracapacitor. Table 9 Characteristics of the energy storage units used in the Saturn Vue test vehicle

Storage unit	Weight (kg)	Volume (l)	Energy stored
Nickel metal hydride battery	24.7	15.4	1.1 kWh (estimated)
Ultracapacitors			
One 48 V module	14.8	11.2	35 Wh
Two 48 V modules	29.6	22.4	70 Wh

driving cycles. Test track acceleration tests of the vehicle indicated no differences in the acceleration times for 0–60 or 40–60 mph using the one module ultracapacitor and nickel metal hydride battery units.

Fuel Cell Vehicle Applications Ultracapacitors can also be used to meet the peak power transients in vehicles powered by fuel cells [22, 23]. The most publicized use of ultracapacitors in fuel cell vehicles has been that by Honda [22] in the “Honda FCV.” In the Honda FCV, the capacitors were used to load level the fuel cell during accelerations and to recovery energy during regenerative braking. In the Honda fuel cell system, the ultracapacitors are used without interface electronics. The capacitors functioned very well and resulted in significant improvements in system energy efficiency.

A detailed study of the use of ultracapacitors in fuel cell-powered vehicles is given in [24]. The study showed that ultracapacitors are well suited to be used with fuel cells. This is especially true when the fuel cell and ultracapacitors are connected in parallel because the effective rest voltage of the capacitor is variable and their high-power capability permits them to meet the maximum power requirement of the system at all voltages between the rest voltage and the minimum voltage at which the fuel cell is to be operated. In this mode of operation, the system is self-controlling with the ultracapacitor and fuel cell sharing the power demand based on the system voltage of operation. Tests [24] were performed on a system (Fig. 5) consisting of the



Energy Storage: Ultracapacitor. Figure 5
Photograph of the ultracapacitor/fuel cell test setup

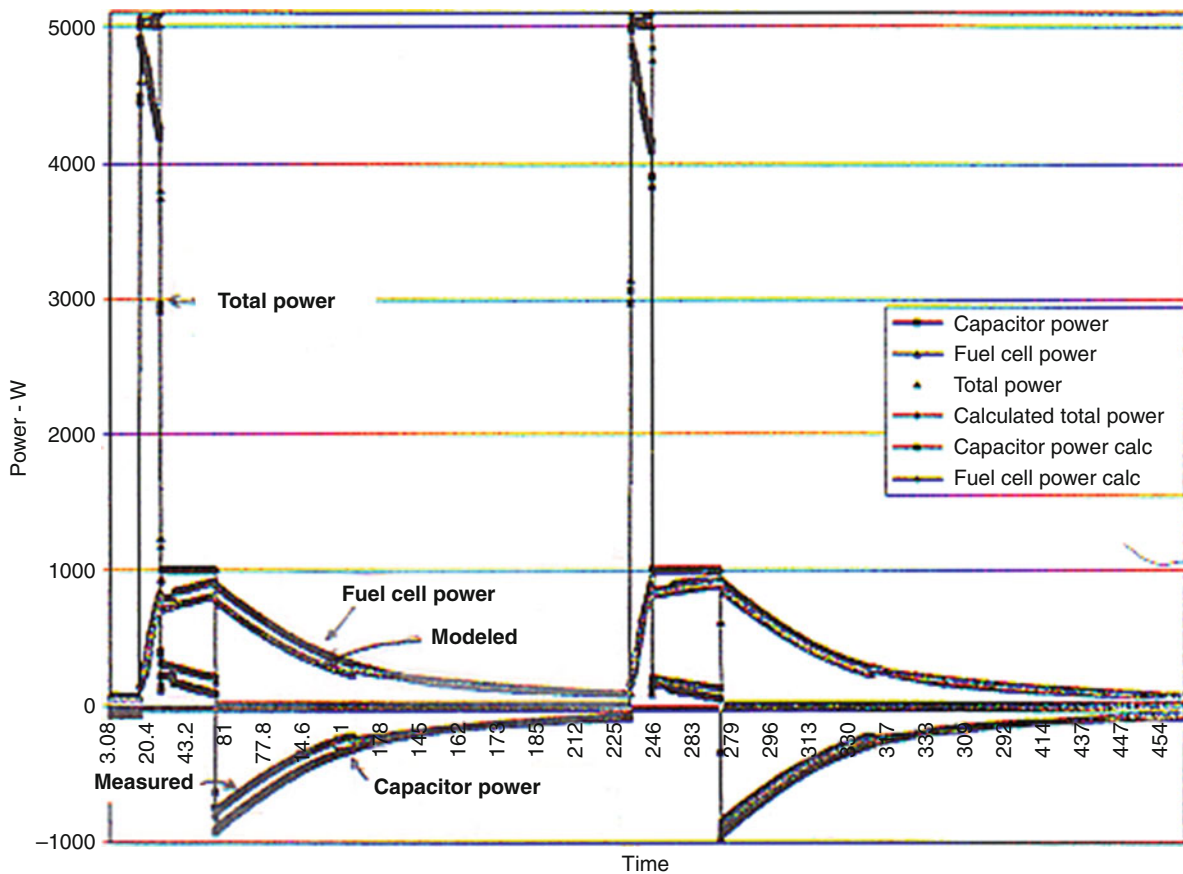
Ballard 1.2 kW Nexa fuel cell and an ultracapacitor bank made up of 17 Ness 2,500 F capacitors. The system was tested for 10 s pulses up to 5 kW. An analytical model of the fuel cell/capacitor system was derived and programmed using an EXCEL spreadsheet. Comparisons of the measured and modeled response of the system at various power levels (Fig. 6) indicated that the system operated as modeled and that the self-control of the fuel cell was as expected. SIMPLER simulations [24] on the FUDS driving cycle were also performed for a midsize SUV powered by ultracapacitors connected in parallel with a fuel cell. It was found that the ultracapacitors significantly load leveled the fuel cell and the behavior of the system was much like that measured with the Ballard 1.2 kW Nexa fuel cell.

Hybrid Vehicle Simulations Using Ultracapacitors Alone

Powertrain Control Strategies with Ultracapacitors Alone in Place of Batteries

Schematics of hybrid powertrains using ultracapacitors and batteries are shown in Fig. 1. If an ultracapacitor unit is used alone, there is no need for special electronics if the motor electronics can handle the increased voltage swing of the ultracapacitors. If that is not the case, then interface electronics is needed between the ultracapacitors and the motor electronics.

When ultracapacitors are used alone as the energy storage unit in a charge-sustaining hybrid (HEV), the objective of the control strategy is to permit the engine to operate near its maximum efficiency. As shown in [25–27], this can be done by operating the hybrid vehicle on the electric drive only when the power demand is less than the power capability of the electric motor; when the vehicle power demand exceeds that of the electric motor, the engine is operated to meet the vehicle power demand plus the power to recharge the ultracapacitor unit. In this mode, the electric machine is used as a generator and the engine operating point is selected along its maximum efficiency line (torque vs. RPM). The recharging power is limited by the power of the electric machine because ultracapacitors can have pulse power efficiency greater than 95% for W/kg values of over 2,000 W/kg (see Table 3). This control strategy is referred to as the “sawtooth” strategy



Energy Storage: Ultracapacitor. Figure 6

Comparison of the measured and modeled ultracapacitor and fuel cell powers for the 5 kW pulse power test

because a plot of the ultracapacitor state-of-charge (SOC) has the form of a saw blade. It can also be referred to as the “high engine efficiency” strategy.

Vehicle Simulation Results

Simulations of midsize passenger cars using the ultracapacitors in micro-hybrid and charge-sustaining hybrid powertrain designs have been performed using the **Advisor** vehicle simulation program modified with special routines at UC Davis [25–27]. All the powertrains were in the same vehicle having the following characteristics: test weight 1,660 kg, $C_d = 0.3$, $A_F = 2.25 \text{ m}^2$, $\text{RRCF} = 0.009$. The engine map used in the simulations was for a Ford Focus 2 l, 4-cylinder engine. The rated engine power was 120 kW for the conventional ICE vehicle and the micro-hybrid and 110 kW for the charge-sustaining hybrids. All the

hybrids use the single-shaft arrangement similar to the Honda Civic hybrid. The same induction electric motor map was used for all the vehicle designs.

The simulation results are summarized in [Table 10](#) for a conventional ICE vehicle and each of the hybrid designs. Results are given for fuel usage in terms of both l/100 km and mpg for various driving cycles. It is clear from [Table 10](#) that large improvements in fuel usage are predicted for all the hybrid powertrains using ultracapacitors for energy storage. The simulation results will be discussed separately for each hybrid design.

Micro-hybrids The results for the micro-hybrids are particularly interesting and somewhat surprising, because of the large fuel economy improvements predicted. These improvements were about 40% on

Energy Storage: Ultracapacitor. Table 10 Summary of the vehicle fuel economy simulation results using ultracapacitors for various driving cycles

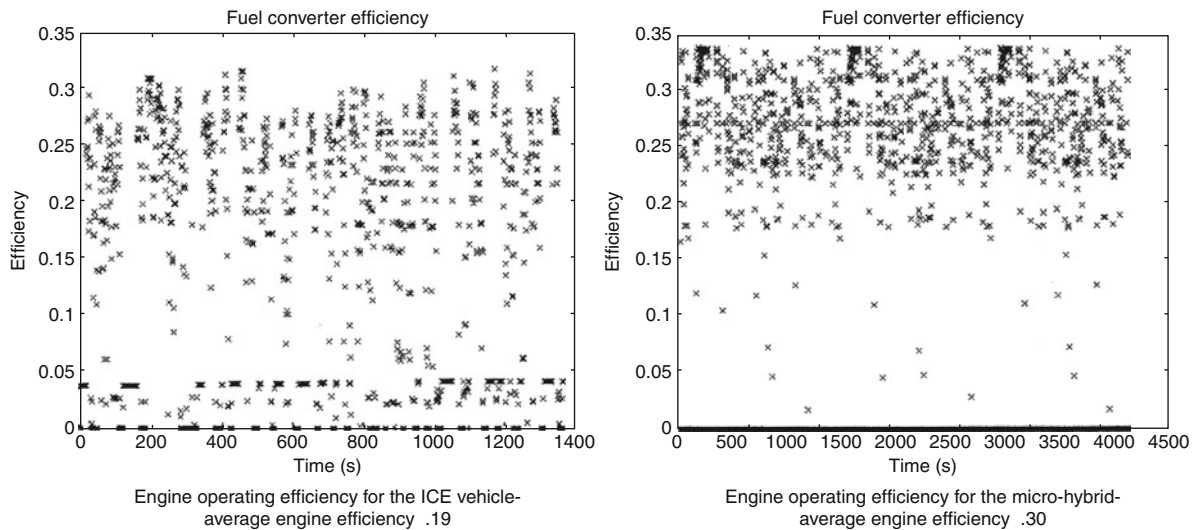
Driveline type	Energy storage type	Voltage and weight cells (kg)	EM Peak kW	L/100 km/mpg			
				FUDS	HWFET	US06	ECE-EUDC
ICE baseline				10/23.8	6.9/34.4	9.6/24.7	9.7/24.6
Micro-HEV	Lead-acid/ultracaps	48					
	Carbon/carbon	6 kg	6	5.7/41.7	5.3/44.7	7.8/30.6	5.9/40.2
	Hybrid carbon	3 kg	6	7.3/32.8	6.3/38.0	8.9/26.7	7.1/33.4
Charge-sustaining hybrid	Ultracaps	200					
	Carbon/carbon	30 kg	35	5.4/43.8	5.0/47.9	7.1/33.6	5.5/43.2
	Hybrid carbon	13 kg	35	5.8/40.9	5.2/45.8	8.0/29.9	5.8/41.3
Plug-in hybrid.	12 kWh Li battery (200 Wh/kg) and ultracaps	300	70 kW with 45 kW from caps				
	Carbon/carbon	40 kg	45	5.5/43.2	5.0/47.7	7.0/33.9	5.5/42.9
	Hybrid carbon	18 kg	45	5.8/41.2	5.2/46.2	7.9/30.2	5.8/41.2

the FUDS and ECE-EUD cycles and 20% on the Federal Highway and US06 cycles using the carbon/carbon ultracapacitors. The improvements are significantly less using the advanced hybrid carbon units because of their lower round-trip efficiencies. In the micro-hybrid designs, the rated engine power used was the same as that in the conventional ICE vehicle in order that the performance of the hybrid vehicle when the energy storage in the ultracapacitors is depleted would be the same as the conventional vehicle. The ultracapacitors were used to improve fuel economy with only a minimal change in vehicle acceleration performance. The control strategy used was the “saw-tooth” strategy discussed in the previous paragraph. As shown in Fig. 7, this resulted in a large improvement in average engine efficiency from 19% in the ICE vehicle to 30% in the micro-hybrid even though the electric motor had a peak power of only 6 kW.

Additional computer simulations were made for higher motor power (up to 12 kW) and larger ultracapacitor energy storage (up to 50 Wh). It was found that the improvements in fuel economy were only marginally greater. However, using a motor power of 3 kW reduced the fuel economy improvement on the FUDS by more than 50%. Note from Table 10

that the fuel economy improvements using the carbon/carbon ultracapacitors were for all the cycles greater than those using the advanced hybrid carbon devices. This was the case because the round-trip efficiencies for the carbon/carbon units were 95–98% and those of the hybrid carbon units were 75–90% for the various driving cycles. As shown in Table 2, the advanced hybrid carbon devices had higher energy density, but even though their power density for 95% efficiency was relatively high (1,050 W/kg), it was not proportionally higher – that is twice as high – as the carbon/carbon devices with lower energy density. These results show that it is essentially to develop high energy density ultracapacitors with proportionally higher power density; otherwise their use in vehicle applications will be compromised.

Charge-Sustaining Mild/Strong Hybrids The fuel economy simulation results for mild/strong charge-sustaining hybrids are also shown in Table 10 for a mid-size passenger car using both carbon/carbon and advanced hybrid carbon ultracapacitors. Using the carbon/carbon ultracapacitor unit, the fuel savings are about 45% for the FUDS and ECE-EUD cycles and about 27% for the Federal Highway and US06 cycles.



Energy Storage: Ultracapacitor. Figure 7

A comparison of engine efficiencies for a conventional ICE vehicle and a micro-hybrid on the FUDS cycle using carbon/carbon ultracapacitor

These improvement values are higher than for the micro-hybrid, but not by as large a factor as might be expected. The prime advantage of the high-power electric driveline in the mild/strong charge-sustaining hybrid is that it yields large fuel economy improvements even for high-power-requirement driving cycles like the US06. The fuel economy improvements using the advanced hybrid carbon ultracapacitor unit are not much less (5–10%) than those with the carbon/carbon unit even though the round-trip efficiency of the hybrid carbon unit is only 85–90% compared to 98% for the carbon/carbon unit. Since the weight/volume of the hybrid carbon unit is relatively small – 43% of that of the carbon/carbon unit, it appears that the mild/strong charge-sustaining hybrid application is a better one for the advanced hybrid carbon technology than the micro-hybrid application.

Plug-in Hybrid Vehicle Simulations Using Advanced Batteries and Ultracapacitors

Control Strategy for Plug-in Hybrid Vehicles (PHEVs)

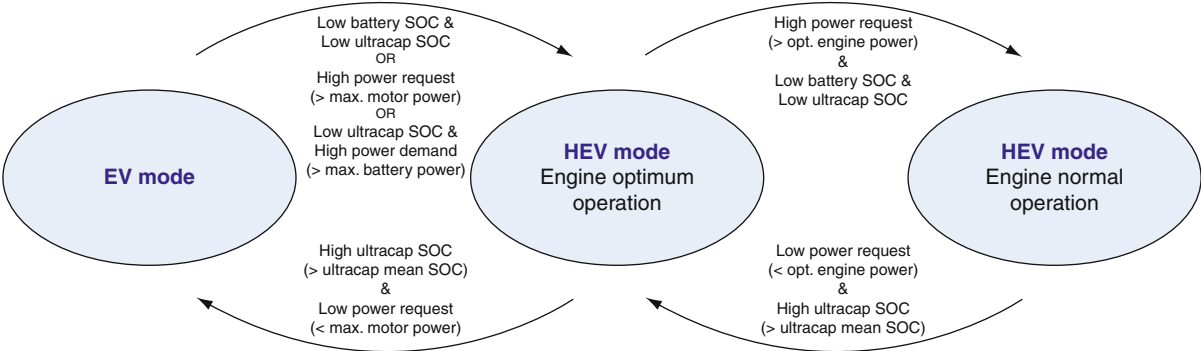
For plug-in hybrid vehicles, batteries and ultracapacitors would be used in combination with

the batteries providing the energy and the ultracapacitors the power. The powertrain schematics are shown in Fig. 1. The control strategy in the charge-depleting mode is to limit the power from the battery to the average power needed by the vehicle with the ultracapacitor providing the additional power during vehicle accelerations. The ultracapacitors also accept all the energy recovered during regenerative braking. If engine operation is needed, the “sawtooth” strategy is used with the ultracapacitors being recharged using engine power. In the charge-sustaining mode of operation of the PHEV, the electric drive is operated using only the ultracapacitors for energy storage like that previously described for mild/strong charge-sustaining hybrid vehicles. The PHEV operating strategies are summarized in Fig. 8.

Simulations of Plug-in Hybrid Vehicles Using Advanced Batteries and Ultracapacitors

A detailed study of plug-in hybrids using advanced batteries is presented in [8]. The characteristics of the advanced batteries used in the simulations are given in Table 11.

The battery and ultracapacitor units used in the simulations are given in Table 12. Simulations were



Energy Storage: Ultracapacitor. Figure 8
Summary of the operating strategy used in the simulation of plug-in hybrid vehicles

Energy Storage: Ultracapacitor. Table 11 Characteristics of present and future battery cell technologies for EVs and PHEVs

Chemistry Anode/cathode	Cell voltage Max/nom.	Ah	Wgt. kg	R mOhm	EV Wh/kg	HEV W/kg 95%	EV W/kg 75%	Cycle life (deep)	Thermal stability
<i>Present technology batteries</i>									
Graphite/NiCoMnO ₂	4.2/3.6	30	0.787	1.5	140	521	2,060	2,000–3,000	fairly stable
Graphite/Mn spinel	4.0/3.6	15	0.424	2.7	127	540	2,120	1,500	fairly stable
<i>Future technology batteries</i>									
Graphite/composite MnO ₂	4/3.6	5	0.09	20	200	250	1,350	–	fairly stable
Silicon carbon composites/composite MnO ₂	4/3.6	20	0.24	4.5	295	621	2,250	–	fairly stable
Rechargeable zinc–air	1.3/1.15	20	60	6.6	385	156	616	–	very stable
<i>Present technology power devices</i>									
supercapacitor activated carbon/activated carbon	2.7/1.35	500 F	0.068	1.3	5.5	2,320	11,600	500 K	Very stable
Power battery lithium titanate oxide	2.8/2.5	4	0.23	1.15	40	1,310	5,170	20–50 K	Very stable

performed with the batteries alone and in combination with the ultracapacitors. The nominal energy storage unit voltage was 240 V in all cases with the maximum currents limited to about 300 A even in the cases of the batteries alone. In all cases, the batteries were depleted to 30% SOC from 100% SOC in the charge-depleting mode of vehicle operation.

Simulation Results for the PHEVs All the PHEV simulations were performed using the following vehicle inputs:
 $C_D = 0.27$, $A_F = 2.2 \text{ m}^2$, $f_r = 0.008$, test weight = 1,650 kg (approx.)
Engine: Honda 1.3 l, iVTEC engine map, scaled to 90 kW

Energy Storage: Ultracapacitor. Table 12 The advanced battery and ultracapacitor units used in the simulations

<i>Lithium-ion batteries</i>				
EIG NiCoMnO ₂	30 kg	20 Ah cells	60 in series	3.5 kWh usable
Composite MnO ₂	32 kg	40 Ah cells	60 in series	4.6 kWh usable
Silicone composite	22 kg	30 Ah cells	60 in series	4.6 kWh usable
<i>Zinc-air batteries</i>				
Zn-air	32 kg	60 Ah cells	180 in series	9.5 kWh usable
<i>Carbon/carbon ultracapacitors</i>				
Symmetric C/C	20 kg	1,350 F cells	110 in series	100 Wh usable

Electric motor: Honda hybrid Civic AC PM 2006 efficiency map, scaled to 70 kW

DC/DC inverter: constant efficiency 0.96

Transmission: 5-speed manual (3.11, 2.11, 1.55, 1.0, 0.71, FD=3.95), automatically shifted in the model, but future models would incorporate the DCT (dual clutch transmission) as a convenient means to have smooth, fast shifting and to decouple the engine when it is not needed to provide power

Fuel and Electricity Use Characteristics The simulation results are summarized in [Tables 13](#) and [14](#) showing results for vehicle operation in the charge-depleting and charge-sustaining modes of the PHEV.

With the batteries in combination with the ultracapacitors, the PHEVs were able to operate in the all-electric mode down to the battery SOC=30% on the FUDS and FDHW driving cycles. In all cases for the US06 driving cycle, the vehicle had blended operation (engine and electric drive both needed) in the charge-depleting mode. The use of the ultracapacitors with the batteries permitted all-electric operation over a wide range of driving conditions with higher Wh/mi for all the driving cycles. Hence in the charge-depleting mode, the fuel economy (mpg) is higher by 50–100% using the ultracapacitors for all the batteries. The fuel economy in the charge-sustaining mode is also higher for all the driving cycles using the ultracapacitors by 15–40%. The acceleration times of

the vehicle were lower using the ultracapacitors than for the batteries alone. With the ultracapacitors, the acceleration times were 2.7 s for 0–30 mph and 6.9 s for 0–60 mph. For the batteries alone, the acceleration times varied somewhat with the battery used ranging from 2.9 to 3.2 s for 0–30 mph and from 8.6 to 9.8 s for 0–60 mph. Hence in all respects, vehicle performance was improved using the ultracapacitors for all the batteries studied.

Reductions in Battery Stress Using Ultracapacitors

The current and voltage responses of the batteries with and without the ultracapacitors are shown in [Figs. 9–12](#) for the silicone carbon lithium-ion and the zinc-air batteries for the FUDS and US06 driving cycles. The effects of the load leveling of the power demand from the batteries using the ultracapacitors are evident in the figures. Both the average currents and the peak currents from the batteries are lower by a factor of 2–3 using the ultracapacitors. The minimum voltages of the batteries are significantly higher using the capacitors and the voltage dynamics (fluctuations) are dramatically reduced. Hence, the stress on the battery and resultant heating are much reduced. The simulation results in [Figs. 9–12](#) also show that the ultracapacitors are utilized over a wide voltage range indicating that a large fraction of their usable energy storage (100 Wh) is being used to load level the batteries. This is only possible using a DC/DC converter between the battery and the DC bus.

The simulation results also indicate that using ultracapacitors, batteries with a wide range of power characteristics can be used in PHEVs and also EVs without sacrificing vehicle performance and subjecting the batteries to high stress and resultant shorter cycle life. This could be especially important in the future as high energy density batteries such as zinc-air and possibly lithium-air are developed. It is likely that those battery types will not have commensurate increases in usable power density and without ultracapacitors, the battery unit in PHEVs and EVs would be sized by the maximum power requirement (kW) rather than the range (mi)/energy requirement (kWh). This would significantly increase weight, volume, and the cost of the battery unit. It is also unlikely that the air electrode will have high charge acceptance

Energy Storage: Ultracapacitor. Table 13 Simulation results for the advanced batteries with ultracapacitors

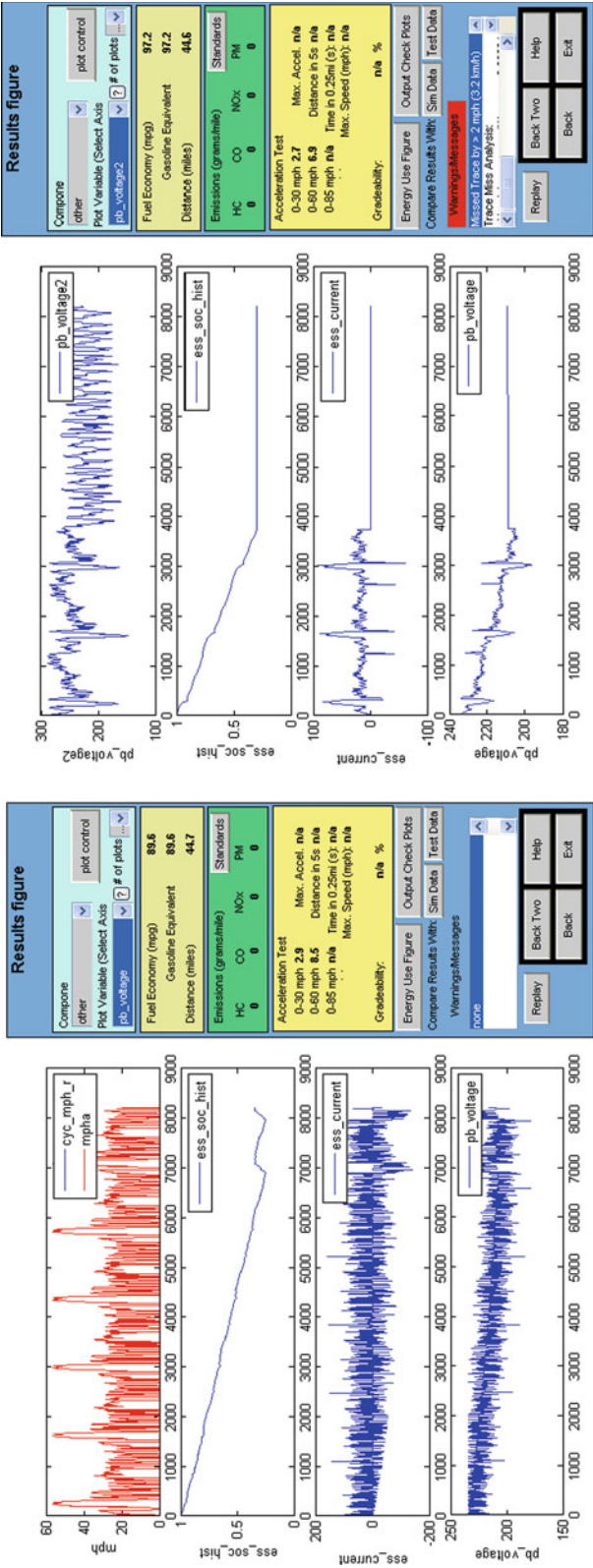
Battery Type (1)	Cycle	Range mi.	kW max. control	kW max bat.	Eff. Bat.	kW max. Cap.	Eff. Cap.	Wh/ mi Bat.	Operat. mode	mpg 20mi	mpg 40mi	Ch. Sus. HEV mpg
Compos. MnO ₂	FUD	22	40	18	0.94	40	0.97	215	AE	none	97	52.8
32 kgbat	HW	20	45	18	0.91	45	0.96	227	AE	none	109	56.3
20 kgcap	US06	30	68	21	0.91	68	0.94	180	blended	71.9	56	38.3
Si Carb/ Compos. MnO ₂	FUD	20	40	18	0.94	40	0.97	220	AE	none	99	52.8
22 kgbat	HW	20	45	19	0.91	45	0.97	225	AE	none	110	56.8
20 kgcap	US06	30	68	21	0.91	68	0.94	190	blended	71.1	52	38.4
Rech. Zn-air	FUD	40	45	19	0.87	45	0.97	228	AE	none	none	54.5
32 kgbat	HW	38	45	19	0.81	45	0.97	242	AE	none	none	57.7
20 kgcap	US06	66	68	21	0.82	68	94	149	blended	62.4	60	38.8

(1) Weight of cells only

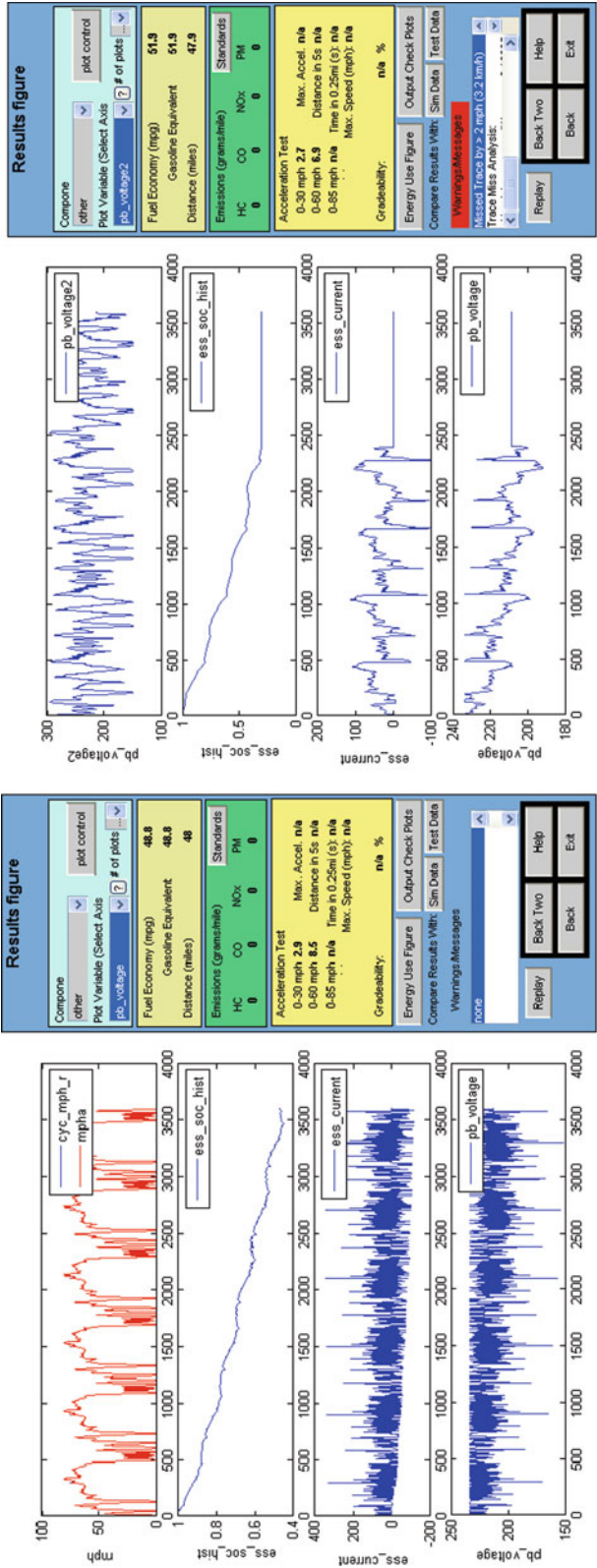
Energy Storage: Ultracapacitor. Table 14 Simulation results for the batteries alone

Battery Type (1)	Cycle	Range mi.	kW max. control	kW max bat.	Eff. Bat.	Wh/mi Bat.	Operat. mode	mpg 20mi.	mpg 40mi	Mpg Ch. sus. HEV
EIG NiCoMn	FUD	27	30	30	0.94	125	blended	134	85	47
30 kg	HW	24	20	20	0.93	137	blended	110	87	47
	US06	57	58	58	0.88		blended	48	45	37
Compos. MnO ₂	FUD	36	30	30	0.92	135	blended	134	104	46.9
32 kgbat	HW	31	20	20	0.91	147	blended	167	113	46.6
	US06	64	58	58	0.87	92	blended	48	48	34.1
Si Carb/Compos. MnO ₂	FUD	35	30	30	0.93	138	blended	138	106	46.9
22 kgbat	HW	32	20	20	0.92	148	blended	169	114	46.9
	US06	64	58	58	0.88	87	blended	48	48	35.7
Rech. Zn-air	FUD	66	30	30	0.84	139	blended	139	137	39.4
32 kgbat	HW	63	20	20	0.83	156	blended	169	169	41.1
	US06	93	36	36	0.72	101	blended	48.5	48.5	30.1

(1) Weight of cells only

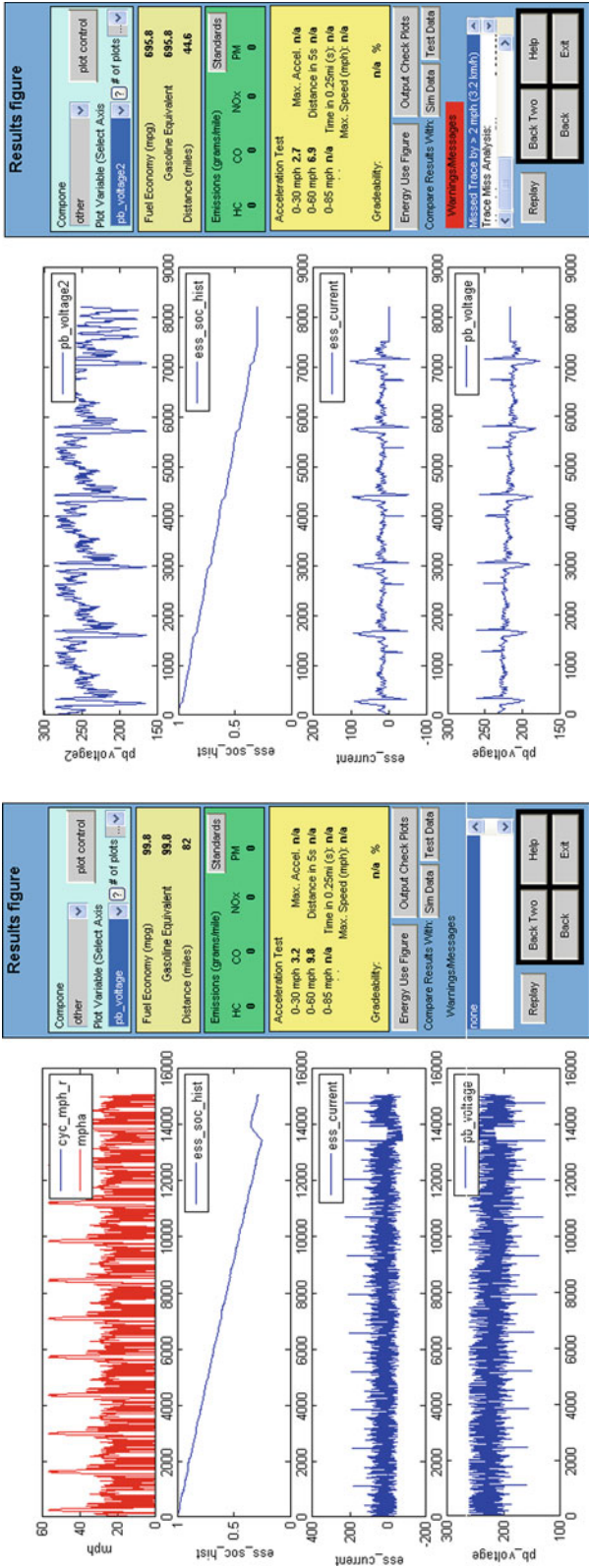


Energy Storage: Ultracapacitor. Figure 9
The Si carbon lithium battery on the FUDS with and without ultracapacitors



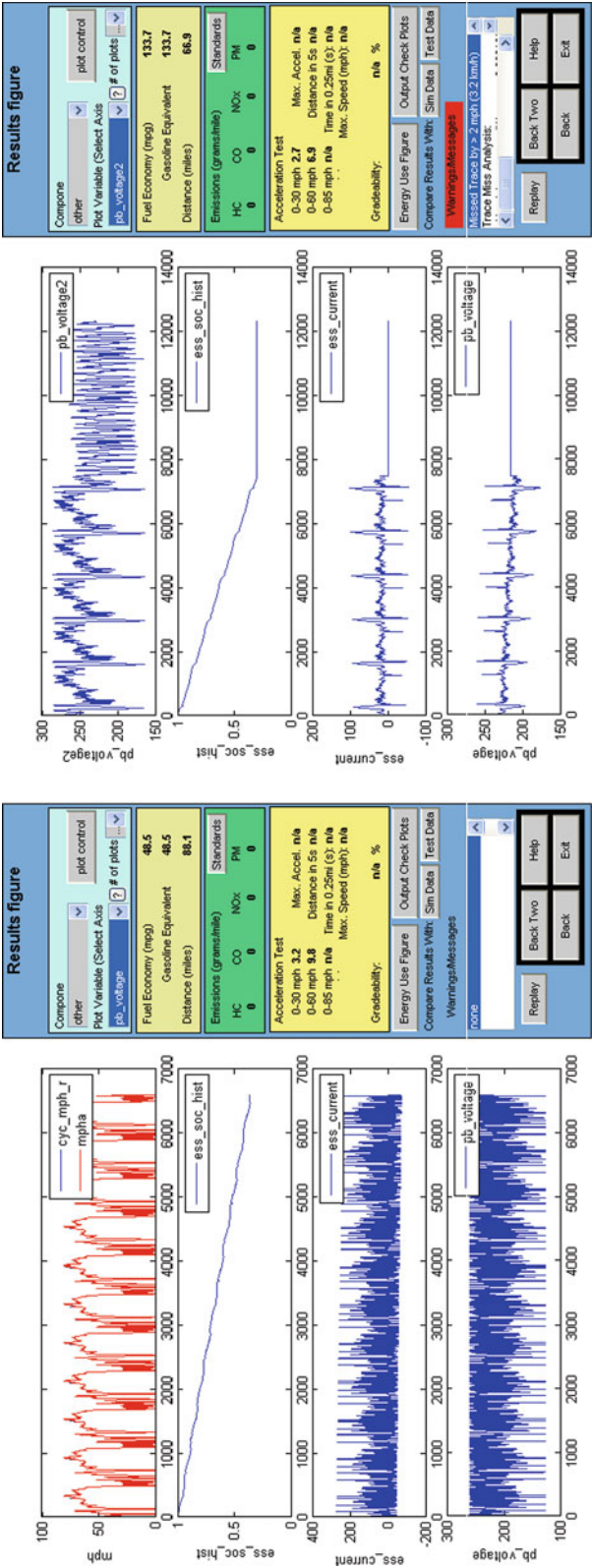
Energy Storage: Ultracapacitor. Figure 10

The Si carbon lithium battery on the US06 with and without ultracapacitors



Energy Storage: Ultracapacitor. Figure 11
The zinc-air battery on the FUDS with and without ultracapacitors

Zinc-air on the FUDS with ultracapacitors



Energy Storage: Ultracapacitor. Figure 12
The zinc-air battery on the US06 with and without ultracapacitors

capability and thus regenerative braking performance approaching that of ultracapacitors or even lithium-ion batteries. This is another advantage of the use of ultracapacitors with the air-electrode batteries.

Fuel Cell Vehicle Simulations Using Ultracapacitors

Fuel Cell Powertrain Configurations and Power Split Strategies

A fuel cell vehicle powertrain consists of three elements: (1) a fuel cell unit that consists of a fuel cell stack, air and hydrogen supply, and water and thermal management systems; (2) an energy storage unit (ultracapacitors or batteries) that can store the electricity generated by the fuel cell as needed; (3) an interface electronics unit that controls the power split between the fuel cell and the energy storage unit. Fuel cells in combination with energy storage can create high power for vehicle traction with fast dynamic response, efficient capture of regenerative braking energy, and reduced stress on the fuel cell due to the elimination of high current pulses. Potentially the use of energy storage can also permit downsizing the fuel cell unit in some applications. As indicated in Fig. 13, there are several approaches to hybridizing the fuel cell system. These various powertrain configurations for fuel cell vehicles have been simulated [28–30] to determine the

effect of the configuration used on vehicle hydrogen consumption and the dynamic stress on the fuel cell stack. Of particular interest in this entry are powertrains using ultracapacitors in place of batteries.

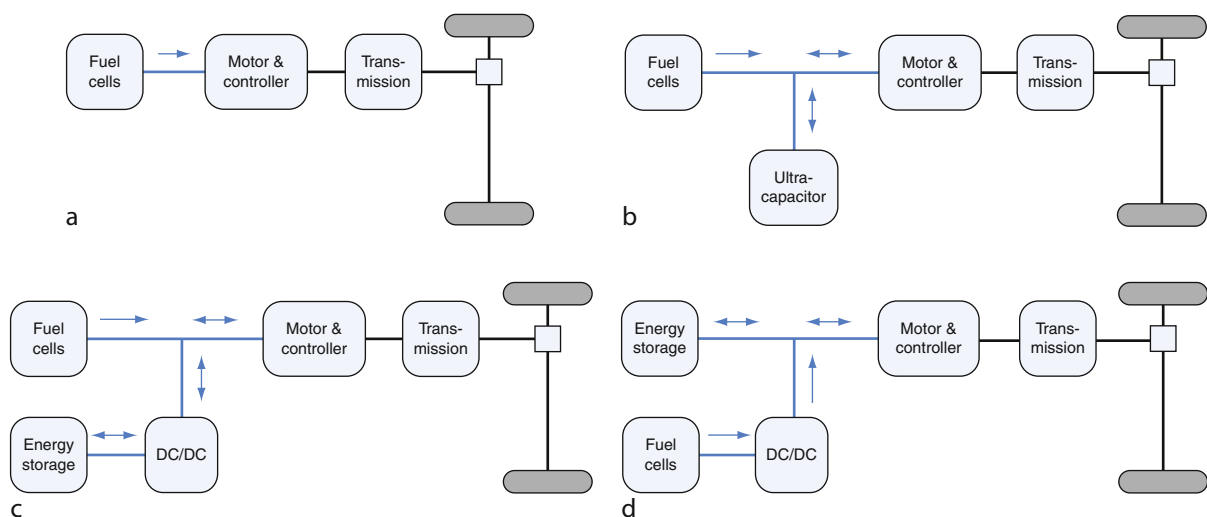
The key considerations in configuring the powertrain are whether interface electronics is used and where the electronics is positioned relative to the fuel cell and the energy storage unit. A powertrain without interface electronics is possible using ultracapacitors, but not with batteries. For fuel cell vehicles with ultracapacitors coupled directly in parallel with the fuel cell, a power diode is employed to prevent the flow of reverse current during braking periods.

In systems without interface electronics, the power sharing between the fuel cell and ultracapacitor is determined by the current–voltage characteristics of fuel cell and ultracapacitor because the voltages must be equal at all times.

$$V_{cap} = V_{fc} - V_{diode}$$

The voltage response of the fuel cell on various driving cycles is modeled in detail in [28–30].

Fuel cell vehicles with DC/DC interface electronics placed between the fuel cell and energy storage units have more flexibility for balancing the power flow between the fuel cell and the energy storage. The general control objective is to operate the fuel cell unit only



Energy Storage: Ultracapacitor. Figure 13
Powertrain configurations for hybridized fuel cells vehicles

in its high efficiency region and as close to constant or slowly varying power as possible. Two different control strategies, power assist and load leveling, can be employed [30]. Both control strategies maintain the SOC of the energy storage unit within a specified range. The *power assist control strategy* sets the current demand for the fuel cell based on the fuel cell voltage and the energy storage SOC. If the fuel cell voltage remains relatively high, it provides most of the current to the electric drive. When the fuel cell voltage becomes low and the energy storage SOC is high, the energy storage unit provides a large fraction of the current. In the *load-leveled control strategy*, the fuel cell provides relatively steady power and the energy storage unit provides transient power. The fuel cell power command is the vehicle demand averaged over a specified time period such as 60 s.

Simulation Results for Fuel Cell Vehicles

The performance and fuel consumption of midsize fuel cell vehicles were simulated for the different powertrain configurations and control strategies using the vehicle models developed at UC Davis [28–30]. Simulations were performed for the hybrid fuel cell vehicles with a dynamic fuel cell system operating in the optimal varying back pressure mode. The simulation parameters are listed in Table 15.

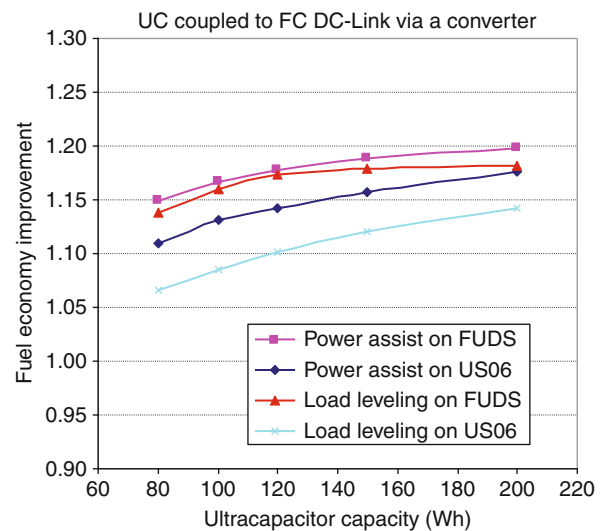
The simulations were performed over the FUDS and US06 driving cycles. The simulation results are discussed in detail in [30]. It was found that the use of energy storage significantly reduced the stress on the fuel cell stack by reducing the peak currents as well as the current fluctuations. There were no significant differences in the stack currents using batteries or ultracapacitors if interface electronics was positioned between the energy storage unit and the fuel cell bus; the results using ultracapacitors without interface electronics were nearly as advantageous as those with electronics.

The use of energy storage in the fuel cell vehicles resulted in improved fuel economy (reduced hydrogen consumption). As shown in Figs. 14 and 15, the improvements increase for larger energy storage units for both ultracapacitors and batteries. In general, the magnitudes of the improvements are much smaller than in hybrid-electric vehicles. This is due to the

Energy Storage: Ultracapacitor. Table 15 Vehicle simulation parameters

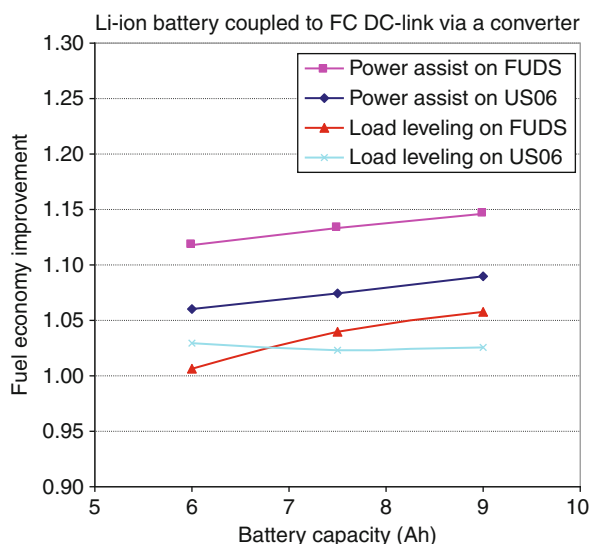
Vehicle and system parameters	
Drag coefficient	0.3
Frontal area (m ²)	2.2
Rolling resistance	0.01
Vehicle hotel load (kW)	0.3
Vehicle mass without energy storage (kg) ^a	1,500
Electric motor (kW)	75
Fuel cell stack and auxiliaries	
Max. net power (kW)	87.6
Gross power (kW)	106
Number of cells	440
Cell area (cm ²)	510
Compressor (kW)	17.2

^aWeight includes all powertrain components except energy storage



Energy Storage: Ultracapacitor. Figure 14 Fuel economy improvements using ultracapacitors

higher efficiency of the fuel cell and its smaller variation with power fraction than is the case for engines. Most of the improvement in fuel economy for fuel cell vehicles using energy storage is due to energy recovery during regenerative braking. Note from Figs. 14



Energy Storage: Ultracapacitor. Figure 15
Fuel economy improvements using lithium batteries

and 15 that the fuel economy improvements with ultracapacitors are slightly larger than with batteries primarily due to the higher efficiency (lower losses) of the ultracapacitors.

Economic Considerations

Reducing the present high cost/price of ultracapacitors is a key issue in achieving high market penetration in the future especially of midsize and large devices. There are many applications for which ultracapacitors are presently precluded or even seriously considered because they remain too expensive even though their selling price has decreased significantly in recent years. The cost of manufacture of any product is closely tied to production volume with the cost decreasing rapidly with increased volume up to relatively high production rates. Sales of capacitors in the many millions of units per year are necessary to reduce the unit costs to levels at which large markets can develop. Semi-automated production facilities presently exist at a number of companies for ultracapacitors of all sizes. In fact, production capabilities exceed sales volumes for most devices and that is the reason the price of devices has decreased markedly in recent years. It is common to speak of the price of devices in terms of cents per Farad (cents/F) or \$/Wh stored. It is easier to interpret the price information on the cents/F basis as it does not

depend on the cell voltage or what fraction of energy stored can be used in a particular application. For example, for a 10 F device, if the price is quoted as 10 cent/F, the device cost would be \$1. Similarly, a 2,500 F device would cost \$25 at 1 cent/F.

The cost to manufacture a carbon/carbon device depends on the material and production costs. At the present time, material costs are high. The cost of carbon suitable for use in ultracapacitors can be as high as \$100/kg with the average price being in the \$30–50/kg range. The cost of the electrolyte solvent is also high in the range of \$5–10/l for both propylene carbonate and acetonitrile. The ionic salts that dissociate in the solvent into the positive and negative ions that move into and out of the double layer in the microporous carbon to store the energy are also expensive being \$50–100/kg. Since the analysis of ultracapacitors is straightforward, material costs can be calculated [31, 32] with good accuracy. The result of a typical costing exercise is shown in Table 16. Note the strong dependency of the cents/F and \$/Wh unit costs for the device on the unit material costs. Presently the price of ultracapacitors is high because both the material and manufacturing costs are high. With more automated production and reduced material costs, it is anticipated that the price of ECCs in high volume can be in the range of 1–2 cents/F for small devices and 0.25–0.5 cents/F for large devices like those needed for vehicle applications.

Ultracapacitors cannot compete with batteries in terms of \$/Wh, but they can compete in terms of \$/kW and \$/unit to satisfy a particular vehicle application. Both energy storage technologies must provide the same power and cycle life and sufficient energy (Wh) for the application. The weight of the battery is usually set by the system power requirement and cycle life and not the minimum energy storage requirement. Satisfying only the minimum energy storage requirement would result in a much smaller, lighter battery than is needed to meet the other requirements. On the other hand, the weight of the ultracapacitor is determined by the minimum energy storage requirement. The power and cycle life requirements are usually easily satisfied. Hence, the unit can be a more optimum solution for many applications and its weight can be less than that of the battery even though its energy density is less than one tenth that of the battery.

Energy Storage: Ultracapacitor. Table 16 Material costs for a 2.7 V, 3,500 F capacitor^a

Carbon			Electrolyte ACN		Device	Unit	Costs		
F/gm	gmC/dev.	\$/kg	\$/l	\$/kg salt	Total mat. \$	\$/kg	\$/Wh	\$/kW	Ct/F
75	187	50	10	125	17.0	29	6.4	29	0.48
120	117	100	10	125	15.5	26	6	26	0.44
75	187	5	2	50	3.6	6.0	1.3	6	0.10
120	117	10	2	50	2.5	4.2	0.93	4.2	0.070

^a4.5 Wh/kg, 1,000 W/kg-95% eff.

Energy Storage: Ultracapacitor. Table 17 Relationships between capacitor and battery costs

Battery cost \$/kWh	Battery cost ^a \$/kW	Ultracap cost cents/F $V_{cap}=2.6$	Ultracap cost cents/F $V_{cap}=3.0$	Ultracap cost ^b \$/kWh $V_{cap}=3.0$	Ultracap cost \$/kW $V_{cap}=3.0$
300	30	0.25	0.34	3,626	7.3
400	40	0.34	0.45	4,800	9.6
500	50	0.42	0.56	5,973	11.9
700	70	0.59	0.78	8,320	16.6
900	90	0.76	1.0	10,667	21.3
1,000	100	0.84	1.12	11,947	23.9

^abattery 100 Wh/kg, 1,000 W/kg

^bcapacitor 5 Wh/kg, 2,500 W/kg

Consider the example of a charge-sustaining hybrid like the Prius. If the energy stored in the capacitor unit is 125 Wh and that in the battery unit is 1,500 Wh, the unit costs of the capacitors and battery for equal unit cost are related by

$$(\$/\text{Wh})_{\text{cap}} = .012(\$/\text{kWh})_{\text{bat}}$$

The corresponding capacitor unit costs in terms of cents/Farad and \$/kWh are given by

$$(\text{cents}/\text{F})_{\text{cap}} = .125 * 10^{-3} * (\$/\text{kWh})_{\text{bat}} * V_{\text{cap}}^2$$

$$(\$/\text{kWh})_{\text{cap}} = 9.6 * 10^4 (\text{cents}/\text{F})_{\text{cap}} / V_r^2$$

and shown in Table 17 for a range of battery costs.

The results shown in Table 17 indicate that for the charge-sustaining hybrid application, ultracapacitor costs of 0.5–1.0 cents/F are competitive with lithium battery costs in the range of \$500–700/kWh. Note also that the \$/kW cost of the ultracapacitor are about one fourth those of the batteries.

Future Directions of Ultracapacitor Development and Vehicle Applications

Ultracapacitor Developments

Ultracapacitors for vehicle applications – that is devices having a capacitance greater than several hundred Farads – have been under development for over 20 years. This development is expected to continue with improvements in energy density as special carbons developed especially for the ultracapacitor application become available. Energy densities of about 10–15 Wh/kg can be expected using the advanced carbons and electrolytes permitting voltages of 3–3.5 V. The cost of ultracapacitors using activated carbon are presently 1–2 cents/F. It is expected that the cost will continue to decrease as the volume of production of ultracapacitors continues to increase. If the cost can be reduced to 0.25–0.50 cents/F, the market for ultracapacitors in vehicle applications will become large (millions of vehicles per year using the devices).

Development of ultracapacitors using pseudo-capacitance or Faradaic materials in at least one electrode will continue. These devices can use either activated carbon or an advanced carbon in one of the electrode. The energy density of these advanced devices could be greater than 20 Wh/kg. The key issues with the advanced devices are their power capability (W/kg for 95% pulse power efficiency) and cycle life. Research to date has indicated it is difficult to achieve large improvements in energy density without significant sacrifices in cycle life and power capability.

Vehicle Applications of Ultracapacitors

There have been numerous demonstrations of the use of ultracapacitors in vehicle applications, but except for a limited number of transit buses and heavy-duty trucks produced for sale, there has been no significant market for ultracapacitors for hybrid vehicles. This may be about to change (2011) as several auto companies are discussing the inclusion of ultracapacitors in micro-hybrids combined with lead-acid batteries. The success of this application is highly dependent on the cost of the ultracapacitors and will require a continuing decrease in their cost. If the energy density of the carbon/carbon devices can be increased to 8–10 Wh/kg with continuing cost reductions to 0.5–1.0 cents/F, it seems likely ultracapacitors will be used in charge-sustaining HEVs to further improve the fuel economy of those vehicles. This could be a very large market for ultracapacitors.

Applications of ultracapacitors in vehicles further in the future could involve their use in fuel cell-powered vehicles and electric and hybrid vehicles using advanced, high energy density (>250 Wh/kg) lithium batteries, which may not have corresponding higher power capability and need ultracapacitors to meet vehicle power requirements. These applications can be the focus of laboratory studies and field demonstrations in the next 5–10 years.

Summary

In this entry, various aspects of the use of ultracapacitors alone and in combination with batteries are discussed in relation to general system considerations and the performance of the energy storage systems in charge-sustaining and plug-in hybrid vehicles and fuel cell-powered

vehicles. The performance of ultracapacitors and lithium batteries is reviewed based on test data taken at the University of California-Davis. It was found that ultracapacitors are commercially available with energy densities of 4.5–5 Wh/kg and power density of 1–2 kW/kg with a pulse efficiency of 95%. Lithium batteries of various chemistries were also tested. It was found that batteries consisting of graphite and manganese oxide/spinel had energy densities of 140–160 Wh/kg and power densities of 0.5–1.0 kW/kg with pulse efficiencies of 95%. High-power lithium batteries using lithium titanate oxide in the negative electrode were found to have an energy density of about 40 Wh/kg and a power density of 1.3 kW/kg with pulse efficiencies of 95%. Comparisons of the power densities of ultracapacitors and lithium batteries indicated that the best ultracapacitors had power densities a factor of 2–3 higher than the high-power batteries, but there are some batteries with higher power capability than the lower power capability capacitors. For use in hybrid vehicles, high power density is very important for ultracapacitors.

The application of ultracapacitors in hybrid-electric vehicles was considered in detail utilizing vehicle simulation results. The use of ultracapacitors alone is analyzed for micro- and mild/full charge-sustaining hybrids and in combination with high-energy density batteries for plug-in hybrids. It was found that ultracapacitors could be used in place of batteries in the charge-sustaining hybrids resulting in fuel economy improvements of 25–50% even for relatively small electric drive systems of 6–10 kW. The ultracapacitor units only stored 50–100 Wh of usable energy and their efficiency was high being 95–98%. In all cases, the ultracapacitors were used with interface electronics that permitted the use of a large fraction of their stored energy.

Vehicle simulations were also performed for plug-in hybrid vehicles that used ultracapacitors combined with advanced, high energy density batteries. The batteries had energy densities of 200–400 Wh/kg, but only moderate power capability of 300–1,000 W/kg for 90% efficient pulses. The simulation results indicated that the advanced batteries combined with ultracapacitors could provide all-electric operation on the FUDS and Federal Highway cycles in the charge-depleting mode and excellent fuel economy in the charge-sustaining mode. Vehicle operation with the advanced

batteries without the ultracapacitors required blended operation on all driving cycles. Comparisons of the current/voltage/power profiles of the batteries with and without the ultracapacitors indicated the peak currents and thus the stress on the batteries were reduced by about a factor of 3 using the ultracapacitors. This reduction is expected to lead to a large increase in battery cycle life.

Vehicle simulation results are also discussed for fuel cell vehicles using ultracapacitors or batteries for energy storage. It was found that the vehicle and fuel cell system functioned essentially the same with ultracapacitor or batteries. In the case of fuel cell vehicles, the ultracapacitor unit could be used without interface electronics with the resultant power split between the ultracapacitors and the fuel cell being self-controlled by the voltage–current characteristics of the respective devices. The fuel economy (hydrogen consumption) improvements using energy storage were 10–15% for the FUDS cycle and 5–10% for the US06 cycle with ultracapacitors and a few percent less using lithium batteries.

The long-term costs of both lithium batteries and ultracapacitors are uncertain. An analysis of ultracapacitor unit costs indicated that a battery cost of \$500/kWh and ultracapacitor cost of 0.5 cents/F would result in a comparable energy storage unit cost for charge-sustaining hybrid vehicles.

Bibliography

- Burke AF, Miller M (2009) Performance characteristics of lithium-ion batteries of various chemistries for plug-in hybrid vehicles, EVS-24, Stavanger (paper on the CD of the meeting)
- Burke AF, Miller M (2009) Electrochemical capacitors as energy storage in hybrid-electric vehicles: present status and future prospects, EVS-24, Stavanger (paper on the CD of the meeting)
- Burke AF, Miller M (2009) The UC Davis emerging lithium battery test project, Report UCD-ITS-RR-09-18
- Burke AF, Miller M (2011) The power capability of ultracapacitors and lithium batteries for electric and hybrid vehicle applications. *J Power Sources* 196(1):514–522
- US Department of Energy (2008) Battery test manual for plug-in hybrid electric vehicles, INL/EXT-07-12536
- Yoshio M, Brodd RJ, Kozawa A (eds) (2009) Lithium-ion Batteries, chapter 12. In: *HEV Applications*. Springer, New York
- Anderman M (2005) Could ultracapacitors become the preferred energy storage device for future vehicles? In: *Proceedings of the 5th international advanced automotive battery conference*, Honolulu, 15–17 June 2005
- Burke A, Miller M (2010) Lithium batteries and ultracapacitors alone and in combination in hybrid vehicles: fuel economy and battery stress reduction advantages. Paper presented at the electric vehicle symposium 25, Shenzhen
- Bohn T (2007) Plug-in hybrid vehicles: decoupling battery load transients with ultracapacitor storage. Paper presented at the advanced capacitor world summit 2007, San Diego
- Miller JM, Bohn T, Deshpande U (2008) DC-DC converter buffered ultracapacitor in active parallel combination with lithium battery for plug-in hybrid electric vehicle energy storage. SAE paper 2008-01-1003, SAE World Congress, Detroit
- Dixon J, Nakashima I, Ortuzar M (2010) Electric vehicle using a combination of ultracapacitors and ZEBRA battery. *IEEE Trans Ind Electron* 57(3):943–949
- Bender J (2009) Life cycle testing of a ultracapacitor/lithium battery combination. Private communication from AFS Trinity
- Bartley T (2005) Ultracapacitor energy storage in heavy-duty hybrid drive update. In: *Proceedings of the advanced capacitor world summit 2005*, San Diego
- King RD et al (2003) Ultracapacitor enhanced zero emissions zinc air electric transit bus – performance test results. In: *Proceedings of the 20th international electric vehicle symposium*, Long Beach
- Okamura M, Hayashi K, Ohta H (2006) Status report on ECasS and Nanogate-capacitors. In: *Proceedings of the 16th international seminar on double-layer capacitors & hybrid energy storage devices*, Deerfield Beach
- Schwake A (2006) EC-funded project “SUPERCAR”: ultracapacitor modules for mild hybrid applications. In: *Proceedings of the second international symposium on large ultracapacitor (EDLC) technology and application*, Baltimore, 16–17 May 2006
- Gonder J, Pesaran A, Lustbader J, Tataria H (2009) Fuel economy and performance of mild hybrids with ultracapacitor simulations and vehicle test results. Presentation at the 5th international symposium on large EC capacitor technology and applications, Long Beach
- Armiroli P et al (2008) Valeo stars+system, ultracapacitors: the way for regen for micro-mild hybrids. In: *Proceedings of the fourth international symposium on large ultracapacitor technology and application*, Tampa
- Peasaran A (2007) Factors and conditions for widespread use of ultracapacitors in automotive applications. In: *Proceedings of the advanced capacitor world summit 2007*, San Diego
- Burke AF (2002) Cost-effective combinations of ultracapacitors and batteries for vehicle applications. In: *Proceedings of the second international advanced battery conference*, Las Vegas, 4–7 Feb 2002
- Knorr R (2006) SUPERCAR – results of a European mild hybrid project. In: *Proceedings of the 6th international advanced automotive battery and ultracapacitor conference*, Baltimore, 17–19 May 2006

22. Knight B (2003) Information on the ultracapacitors used in the Honda fuel cell vehicle
23. Thounthong P, Rael S, Davat B (2009) Energy management of fuel cell/supercapacitor hybrid power source for vehicle applications. *J Power Sources* 193:376–385
24. Burke AF, Miller M (2003) Ultracapacitor and fuel cell applications. In: *Proceedings of the 20th international electric vehicle symposium*, Long Beach
25. Burke AF, Zhao H, Van Gelder E (2009) Simulated performance of alternative hybrid-electric powertrains in vehicles on various driving cycles, EVS-24, Stavanger, (paper on the CD of the meeting)
26. Burke AF, Van Gelder E (2008) Plug-in hybrid-electric vehicle powertrain design and control strategy options and simulation results with lithium-ion batteries. Paper presented at EET-2008 European Ele-drive conference, Geneva, 12 Mar 2008 (paper on CD of proceedings)
27. Burke AF (2007) Batteries and ultracapacitors for electric, hybrid, and fuel cell vehicles, *IEEE Journal*, Special issue on Electric Powertrains
28. Zhao H, Burke AF (2009) Optimum performance of direct hydrogen hybrid fuel cell vehicles, EVS-24, Stavanger (paper on the CD of the meeting)
29. Zhao H, Burke AF (2009) Optimization of fuel cell system operating conditions for fuel cell vehicles. *J Power Sources* 186(2):408–416
30. Zhao H, Burke AF (2010) Effects of powertrain configurations and control strategies on fuel economy of fuel cell vehicles. Paper presented at the electric vehicle symposium 25, Shenzhen
31. Burke AF (2010) Electrochemical capacitors. In: *Handbook of batteries*, 4th edn. McGraw-Hill, New York
32. Burke AF (2004) Key issues for the implementation and marketing of ultracapacitors in vehicle applications. In: *Proceedings of the 14th international seminar on double-layer capacitors and similar energy storage devices*, Deerfield Beach

Engineered Geothermal Systems, Development and Sustainability of

ROY BARIA¹, L. MORTIMER², G. BEARDSMORE²

¹EGS Energy Ltd, Penance Cornwall, UK

²Hot Dry Rocks Pty Ltd, South Yarra, Victoria, Australia

Article Outline

Glossary

Definition of Engineered Geothermal Systems

Introduction

Future Directions

Acknowledgments

Bibliography

Glossary

Breakthrough time The time taken for a tracing element (chemicals) to travel through a specified media (a reservoir) indicating a degree of channeling or dispersion of injected fluid passing through the reservoir.

EGS Enhanced or Engineered Geothermal system; a geothermal system, a geothermal system engineered to mimic a natural system (hydrothermal) so that energy can be extracted from hot rocks at great depth.

Granite Granite is a common and widely occurring type of intrusive, felsic, igneous rock. Granites usually have a medium- to coarse-grained texture and can contain radiogenic material which produces heat in the rock.

Igneous rocks Igneous rock is formed through the cooling and solidification of magma or lava. Igneous rock may form with or without crystallization, either below the surface as intrusive (plutonic) rocks or on the surface as extrusive (volcanic) rocks.

In situ stress Forces that exists in the crustal plate and the three mutually orthogonal components are categorized as vertical, minimum horizontal, and maximum horizontal. It plays a very important part in the development of an EGS reservoir.

Joint network Term used to collectively describe fractures within a rock mass characterised by features such as fracture orientation, density, spacing, length and mechanical stiffness.

Permeability Permeability is defined by as a measure of flow through spacing between the two faces of a fracture open enough to allow fluid flow through it.

Reservoir A matrix of fractures are manipulated by injecting fluid under pressure for fractures to slide against each other, leaving a permanent increased aperture in the fracture so that fluid can pass through it to recover heat.

Shear failure A terminology used when two sides of a fracture slip over each other with minor permanent displacement.

Tensile failure A terminology used to create a new fracture by splitting the rock mass using fluid injection. The fracture remains open while the pressure is applied but closes once the pressure is taken off, leaving no residual gap between the two new faces of the fracture.

Thermal drawdown A process whereby the reservoir is progressively cooled over time by the passing through of cooler re-injected fluid.

Definition of Engineered Geothermal Systems

Engineered Geothermal System (EGS) is a relatively new technology with tremendous potential for providing heat and power, as well as helping to address the issue of reduction of CO₂ in the environment. The technology is complex, and it has taken some time for a series of research projects globally to understand the physical processes involved, to develop supporting technologies such as high temperature instrumentation, numerical models, and to validate the concept. In view of EGS being a new technology with very limited operating history, the issue of sustainability is difficult to address. It is important to assess the sustainability issue based on the current state of knowledge whilst also recognizing the fact that some existing data are inconclusive. Observations and conclusions reached at many of the EGS sites worldwide have been incorporated to give a broad-based qualitative assessment and to reduce the bias.

The most advanced and near commercial-scale EGS project is the European project at Soultz-sous-Forêts, north of Strasbourg in France. Knowledge gained from over 30 years of research carried out at other EGS projects formed the basis upon which the European project was built [1, 2]. To a large extent, the preceding national EGS projects were at a level of university research and subsequently remained starved of funds. This made it difficult to demonstrate the true potential of EGS. Establishing a European Project [3, 4] provided a big boost and allowed work to proceed on a realistic scale to address the technical and economic targets. The first successful circulation of an EGS system was achieved at Soultz in 1997, at a depth of around 3,600 m, and indicated that technically and economically the technology was promising [5]. Data from Soultz are drawn upon in this paper to support the case why this technology needs to move to a commercialization

scenario, and how it can be scaled up to produce a large, sustainable power output to make it acceptable as a serious contender as an energy provider. It is important to re-emphasize that discussions and arguments for the sustainability of EGS presently remain theoretical as there are very few operating plants at present. However, our judgement and experience has been used to put forward sustainability scenarios based on the history of the development of this technology and observations to date.

Introduction

Global demand for energy is expected to continue to rise, but on the timescales of decades to centuries, the supply of traditional energy resources derived from hydrocarbon are unlikely to keep pace with the increasing demand. Additionally, there is increasing awareness of global warming, the political dimension to the generation of CO₂, and the importance of domestic energy security. Nations, industry, and the public are aware that low-emission energy resources will have to play an increasingly important role as people both choose and are forced to move away from hydrocarbon-based energy resources. The effect of digressions between the increasing demand and the limited supply of hydrocarbon-based energy will manifest itself as an increasing price of energy. The increasing prices should, in principle, make some renewables cost competitive.

The energy potentially available from Geothermal resources generally, and EGS in particular, is huge; almost beyond comprehension. Theoretical calculations indicate that cooling just 1 km³ volume of rock from 200°C to 180°C (i.e., by 20°C) could provide enough thermal energy to power an electric generator of 10 MWe capacity for 20 years, the equivalent of 1.3 million tons of oil. For example, on a slightly larger scale (national scale), the estimate of the Technical Potential for EGS in Australia is around 6,800 GWe [6], which equates to 136 times the total current installed capacity in Australia from all energy sources.

Basic Concept of EGS

The concept of extracting heat energy from the deep, hot rock mass in the earth using EGS technology is relatively simple [7]. Drilling to an appropriate depth can provide access to any desired temperature, and

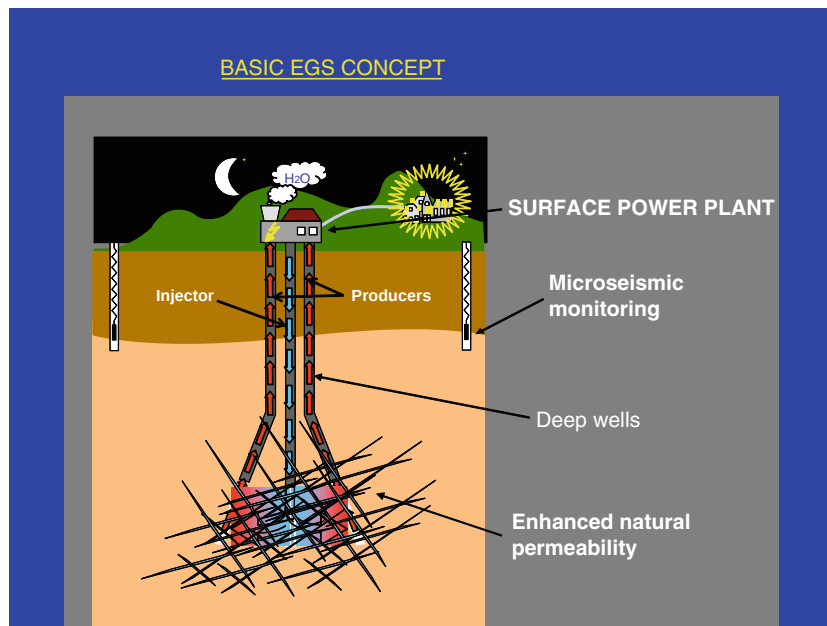
a reservoir (heat exchanger) can be created by pumping cold fluid at high pressure and flow rate from the surface to open up existing fractures, a process known as “reservoir stimulation”. These fractures remain open permanently, giving the stimulated rock mass an enhanced permeability. A second, and sometimes a third, well is drilled into the heat exchanger some distance (typically around 600 m) from the first well. Cold water is pumped down the injection well, the water flows through the enhanced permeable zone (heat exchanger) where it heats up, and the hot water is returned to the surface via a production well in order to drive a turbine to produce electricity and/or space heating. The cooled fluid is then pumped down again and reheated, thus forming a continuous closed circuit loop (Fig. 1).

Although this was the initial concept of EGS, the term “EGS” now also encompasses the exploitation of fluid flow at depth in faults with favorable orientation within the prevailing stress field [8, 9]. Experience and knowledge to date indicate that some large faults that are approximately aligned in the maximum horizontal stress direction at depth are likely to be open and are able to deliver large flow rates of hot fluid for power generation [10]. A production well is drilled to

orthogonally intersect at depth a large fault that is striking in the direction of maximum horizontal stress. Significant flow rates might be immediately achievable. If not, then stimulations may be necessary to enhance the flow rate by decreasing the flow impedance. A similar technique is used to drill a second well (injection well) some distance away and offset from the first fault. In planning such a strategy, it is important to recognize that stimulation of either or both wells may be necessary to decrease the natural impedance. An EGS project based on this model has already been commercialized in Landau, Germany (www.geox-gmbh.de/en). The EGS reservoir has been producing for around 3 years with no indication of reduction in the flow rate or thermal drawdown. The data from the project is not widely circulated due to the commercial nature of the project.

Numerical modeling of the stimulated deep reservoir at Soultz reveals that a reservoir of such dimensions should last for around 20+ years [11] before thermal drawdown reducing the rock temperature below commercial levels. But the results of the numerical models are yet to be confirmed through long-term circulation.

Any, even superficial, estimate of global EGS potential reveals a huge store of thermal energy at relatively shallow



Engineered Geothermal Systems, Development and Sustainability of. Figure 1
Basic EGS concept

depths in the Earth's crust. Debate often surrounds issues such as whether the depth for estimating EGS potential should be limited to 5,000 or 10,000 m. A seminal report from the Massachusetts Institute of Technology in the USA [12] supported 10,000 m as the maximum depth of exploitation for EGS with a view that the technology to exploit energy at this depth is within the grasp of the hydrocarbon industry and therefore could be adapted to work for EGS in the future.

Global Potential of EGS

EGS technology has advanced significantly in the last 20 years. Some of the main questions now raised by potential investors are:

- What is the worldwide potential for EGS?
- Can present technology be scaled up to produce significant amounts of power?
- What are the risks associated with this technology?
- Are EGS reservoirs sustainable?

Taking notice of the concern of potential investors, one of the tasks under the auspices of "Geothermal Implementing Agreement" of the International Energy Agency (IEA/GIA) has been to assess the global potential for EGS. To this end, the Google organization has sponsored a mechanism to quantify the global potential by developing a Protocol, with the intention that it be adopted internationally [13]. Other questions are addressed later on in the paper.

Brief History of the Development of EGS

EGS projects have been activated in many countries, but only four are mentioned here to summarize the development of this technology.

The concept of extracting heat from hot, but low permeability, rocks at great depth and converting this to electricity came from Los Alamos, USA, in the early 1970s. The concept was called Hot Dry Rock (HDR), and a research project was established in a caldera near Los Alamos, New Mexico, USA [14, 15]. The concept relied on some basic assumptions at the time; primarily, that the system had to be in an igneous basement rock with low permeability so that it could be stimulated using high-pressure fluid injection to enhance permeability with minimal chance of losing water in the far field during circulation. The method of enhancing the

permeability of the potential heat exchanger was borrowed from the hydrocarbon industry. It was felt that this technique could, in principle, be replicated worldwide to provide power and heat wherever necessary.

Although the concept appeared to be relatively simple, extensive supporting technologies had to be developed to overcome difficulties such as high temperatures encountered during drilling, casing and cementing, characterization of the development of the reservoir, and establishing the physics associated with it [2].

In the mid-1970s, a research project was established in the UK at the Rosemanowes site in Cornwall [16] to focus on and characterize the physics associated with the creation of an underground heat exchanger. The project targeted a much shallower depth and lower temperature than the Los Alamos site. This was done in recognition of the fact that the high temperature and complex stresses regime had created insurmountable difficulties for Los Alamos. The project at the Rosemanowes site was designed so that the physical processes involved in the development of an EGS reservoir could be easily observed and understood.

The Rosemanowes project addressed many of the uncertainties in the physical processes involved in the creation and characterization of an EGS reservoir, and in particular the influence of the prevailing tectonic stress regime on the fluid flow and water losses [17–19]. A 5-year-long circulation test was also carried out (the longest circulation so far in any EGS project), stimulations were carried out using various materials (explosives, gel with various viscosities, water, etc.). Attempts to define a number of economic targets for assessing the commercialization potential of the technology were also carried out [20].

By the early 1980s, as the definition and the concept of EGS grew broader, it was recognized that "*An EGS system is any system where reinjection is necessary to extract the heat at a commercial rate for a prolonged period*" [21]. If one accepts this definition, then a large proportion of hydrothermal systems may fall within the realms of EGS. Additionally, various names were given to the same concept (Hot Dry Rock) for political acceptance and because of variations in the in situ rock characteristics. Alternative names include "Enhanced Geothermal Systems," "Hot Fractured Rocks," and "Hot Wet Rocks," but, recently, all of these have been

grouped into a more coherent title, called *Engineered Geothermal Systems*. This recognizes the fact that there is not a single solution for creating a reservoir at depth due to the natural variation in rock properties, and that an engineered solution will need to be found to address local conditions based on past experience.

The Soultz project was established in late 1988 about 50 km north of Strasbourg in Alsace, France, near the German border [3, 22, 23]. The project was established as a European project, supported by the European Commission, incorporating Britain, France, and Germany (with Britain leaving the program around 1993). The program at Soultz had an international dimension and sufficient funds to take EGS technology to the next step in development, i.e., bring EGS closer to a commercial prospective by addressing/improving economic targets. The project successfully addressed some of the economic targets by initially creating an EGS reservoir at 3.5 km depth [24], and later at 5,000 m depth. Its international status (Fig. 2) helped to accumulate the knowledge gained in earlier projects and bring it into focus to address the further development required for commercial exploitation [24–28].

Economic Targets for Commercial Acceptance

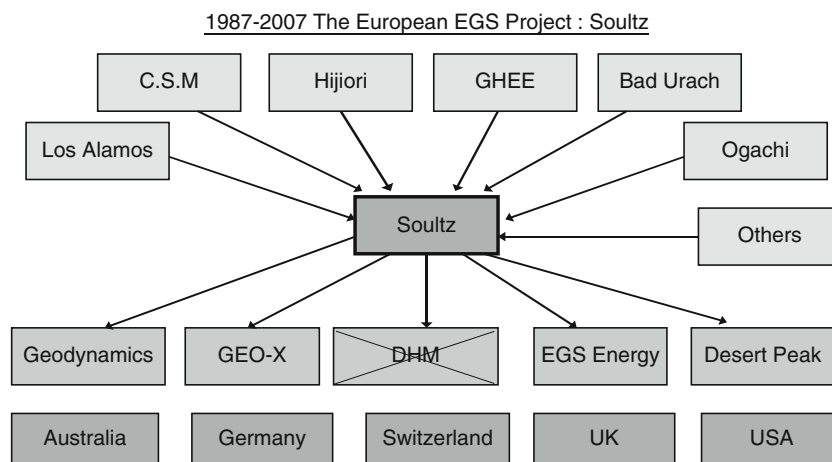
A number of economics models have been generated in the past to assess the parameters necessary to make EGS

commercially viable. One such economic model was prepared by Shock [29] for the UK DoE. That study identified a number of parameters (Fig. 3) that need to be demonstrated and quantified in real terms to gain confidence in a created EGS reservoir. These are:

Quantity of Power Output: 5–6 MWe This depends on the combination of output temperature, the delivered flow rate, and the density of the fluid. For example, at 200°C, one would require a flow rate of ~75 l/s using fresh water. The highest temperature accessed to date in an EGS project is 278°C at the Geodynamics site in Australia [30]. The highest flow rate produced to date is ~100 l/s at the Landau site, Germany.

Life of an EGS Reservoir at the Above Flow Rate In principle, a reservoir is designed to last around 20 years with a temperature drawdown of 10%. For example, if the initial temperature at the inlet of a power plant is 200°C, a temperature drawdown of 10% will bring this to 180°C. Once the system reaches around 180°C, the plant will be less efficient and is likely to be shut down, but this does not mean that there is no heat energy left in the rock mass.

For heat to be extracted efficiently, cold injected fluid has to come into contact with the hot rock so that heat can be transferred from the rock to the fluid. Numerical models suggest that about 1 million square



Engineered Geothermal Systems, Development and Sustainability of. Figure 2

Geothermal Systems, Engineered or Enhanced (EGS), Accumulation and dissemination of knowledge at the European EGS project at Soultz

ECONOMIC REQUIREMENTS	
DRIVEN BY ECONOMICS:	Target 5–6 MWe/module
• LIFE OF THE SYSTEM:	~20 Years
• TEMP/DEPTH OF THE WELLS:	~200 C
• SEPARATION BETWEEN WELLS:	~600 m
• PRODUCTION FLOW RATE:	~75 Kg/s
• FLOW IMPEDANCE:	0.1 MPa/l/s
• WATER LOSS:	~10% MAX
• THERMAL DRAWDOWN:	~10%
• CONTACT SURFACE AREA:	~10 million m ²
• RESERVOIR ROCK VOLUME:	~300 million m ³
• INTEREST RATE FOR THE CAPITAL:	~5%
• SUPPORT:	No CO ₂ levy support etc
Economic study by Shock ~1986 for UK DoE	

Engineered Geothermal Systems, Development and Sustainability of. Figure 3
Main economic targets for EGS (From [29])

meters of contact surface area is required between the fluid and the rock mass [12]. For an EGS reservoir to last 20 years, a numerical model indicates that a total rock volume of around 200 million cubic meters is required [12]. Numerical modeling also indicates that this will necessitate a separation between the wells (injection and production) of around 600 m to achieve the above rock volume.

Parasitic Losses These are defined as a reduction in the net profit caused by the use of power required by various operations. Some of these are associated with the power required for pumping the fluid around the system and is quantified in MPa/L/s. Economic models suggest that 0.1 MPa/l/s is acceptable. Additionally, the system may lose some of the injected fluid in the far field due to the character of the site. In this case, additional water may be required to keep the system pressurized. This is normally quoted in a % of the exit flow. This additional water has to be paid for and comes out of the net profit. This is incorporated in the economic model, and around 10% is included as acceptable losses, but the lesser the amount of fluid losses, the better will be the return on the investment. Recent development of EGS reservoirs shows that typical water losses would be around 1–3%.

One of the key figures in the exploitation of the technology is the interest rate on the borrowed capital. This can play an important part in the net return on the investment as the construction of the overall plant may take anything from 3 to 4 years, depending on the depth, number of wells, etc.

More recently [31, 32], revisited economic modeling of a “Soultz-like” project. One of their key findings was that parasitic pumping power significantly impacts on the net power output from an EGS plant (see also section Parasitic Losses below). Higher flow rates require greater pumping power, and there will typically be an optimal pumping/flow rate to maximize the economic return of the system. The optimal pumping/flow rate might not necessarily be the maximum achievable flow rate.

Technical and Economic Achievements to Date

Reservoir Creation Mechanism The initial focus was on applying the hydrofrac technique used in the hydrocarbon industry for creating an EGS reservoir. This consisted of creating a penny-shaped single new fracture (tensile) by injecting fluid under high pressure [33–35]. Subsequently, it was recognized that this was not the way forward as a penny-shaped single fracture

does not create sufficient heat transfer area nor the fractured volume required for sustained heat extraction at high temperature for prolonged period.

Observations during the creation of EGS reservoirs showed that the reservoir in a natural-jointed rock mass grows by shear failure on existing fractures [17, 18]. Shear failure occurs when the pore pressure in the fracture increases and reaches the value of normal stress on the fracture, causing the fracture surfaces to part and for shear failure to take place. After shear failure, joints are propped open by asperities. Observations before and after stimulation at Soultz demonstrated an increase in “productivity index” of wells GPK2 and GPK4 by a factor of ~ 20 [36]. This, in turn, effectively reduces the pumping power required to circulate fluid through a system, as well as providing increased heat transfer area and fractured rock volume. Additionally, the shear mechanism produces microseismic noise, which is used to monitor the influence of stimulation in real time [18].

Size and Lifespan of the Reservoir The separation between the injection well and the production well, to a large extent, determines the size of the reservoir. Initially, this was in the range of 120–180 m because of a concern that hydrofracturing in granite would be difficult and that it might not be possible to link two wells hydraulically over many hundreds of meter. Following the change in the concept of reservoir creation mechanism (from predominant tensile to shear failure), a separation in excess of 600 m between the injection and production well is regarded as normal. The largest separation of wells at Soultz is 670 m [37–40], and over 1,000 m at the EGS project in Landau, Germany. Numerical modeling indicates that with a separation of this order, an EGS reservoir ought to last over 20 years [41] and that using three wells instead of two (i.e., one injector and two producers) may extend the life even longer. However, no EGS reservoir has yet been circulated this long for confirmation.

One of the methods used for assessing the size of an EGS reservoir is by using tracers [42]. Breakthrough time and model volume are good indicators.

At Soultz, it was felt that the required heat transfer and rock volume were achieved between GPK2 and GPK3, but not between GPK3 and GPK4. The poor

connection between GPK3 and GPK4 was due to a poor stimulation strategy [38, 39], further explanation for this is given in [section Parasitic Losses](#).

Parasitic Losses This was a big issue at the early stages of the development of EGS technology. One of the main energy losses is due to the power required to pump the fluid through the reservoir to collect the heat energy. This resistance to flow (impedance) is quantified as pressure required to flow per liter per second (MPa/l/s). In early days, until 1990, the best impedance achieved was around 0.4 MPa/l/s with a flow rate of 15 L/s with a separation between wells of around 200 m at the Rosemanowes site in the UK. At Soultz, this was reduced to ~ 0.23 MPa/l/s with a separation of 450 m between wells and a production flow of ~ 26 l/s. This EGS reservoir was at 3.6 km depth. It was felt that the impedance and flow rate could be improved further by having a three well system (one injector and two producers). Subsequently, three more wells (GPK2, GPK3 and GPK4) were drilled to 5,000 m depth to access $\sim 200^\circ\text{C}$ rocks, with a separation in excess of 600 m. Following stimulations of GPK2/GPK3, the flow impedance was found to be around 0.29 MPa/l/s, thus for three wells one would expect to half this to around 0.15 MPa/l/s [37, 38]. GPK4 was also stimulated, but a reduction in impedance was not achieved because GPK4 was poorly connected to GPK3, most likely, due to a change in the strategy for stimulation [38].

Stimulation between GPK2 and GPK3 was carried out using a newly developed stimulation technique [38] that involved injecting in both wells simultaneously (“focused stimulation”). GPK4 was stimulated by injection in one well only. Additionally, there was an indication that the previous stimulation in GPK3 may have created a stress barrier between GPK3 and GPK4. High pressures would have been required to break through this barrier, which might have been possible using focused stimulation [38] but was not possible by stimulating just one well.

Tracer tests indicated a breakthrough time between GPK3 and GPK2 in the order of 3.5 days, and between GPK3 and GPK4 of 28 days [42].

Power Output In view of the poor hydraulic connection between the wells GPK3 and GPK4 at Soultz, the flow rate achieved was lower than expected and

therefore the full potential of the system was not realized. Originally, it was planned to produce around 30–35 l/s from each production well, giving a total flow of around 60–70 l/s and power generation of 3–4 MWe. This proved impossible due to poor connection between GPK3 and GPK4. The plan at present is to produce around 2 MWe and gradually develop further to increase the flow rate and characterize the existing reservoir during the circulation period [36].

On the other hand, the EGS project at Landau increased the flow rate by stimulation from 50 to ~100 l/s, and power production of between 2.8 and 3.5 MWe has been achieved (personal communication).

Future Scenarios Which May Lead to a Mature EGS Technology

EGS has distinct advantages over other renewable energy sources in that it can deliver a base load or a peak load, it is a strategic (domestic) resource, has low visual impact, and has a very low carbon foot print (Fig. 4) in the range of 10 g CO₂/kWh.

Some of the disadvantages are finding the right environment for EGS development, the high cost of developing this technology, uncertainty about underground rock properties (although this is dramatically reduced after the drilling of the first well and obtaining the in situ properties), and community concern over

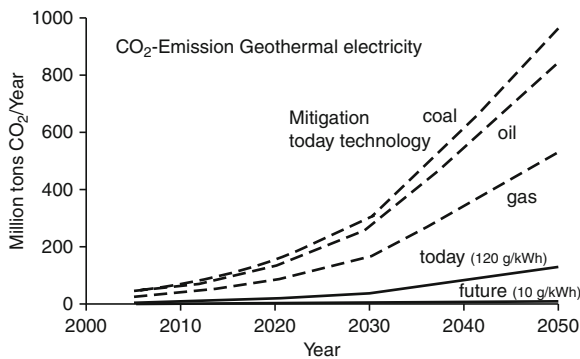
generating seismic events (although experience suggests that this can be managed and mitigated). For example, the EGS project in Basel, Switzerland (DHM), generated a seismic event of the magnitude 3.4 while injecting to create a reservoir. The consequence was the shutdown of the project due to complaints of the local residents and consequence associated with it. This could have been avoided by taking the guide line provided by the protocol on induced seismicity [43] and a white paper published by IEA/GIA [44]. One of the recommendations in the protocol is to avoid establishing an EGS project in the area where there has been large earthquakes in the past or such continuous activities. In 1356, the city of Basel was almost entirely destroyed by a 6.5-magnitude earthquake. Basel is a known moderately active earthquake zone in Switzerland [45] and a possible way forward would have been to explain the situation to local residents, seek their agreement to go forward, and limit the stimulation to smaller volume.

Overall, the advantages of EGS appear to far outweigh the disadvantages, providing a power stimulus for continued development of the technology. Uptake will be progressive, starting with the technically and economically easiest styles of development and progressing to more challenging environments.

It is becoming apparent for the initial rapid take-up of this technology the easiest targets are large faults with high permeability, as at Landau, Germany. Observations at Soultz and other places have indicated that large faults often exist at great depth in granite basement, and those that are roughly aligned with the maximum horizontal stress orientation can potentially deliver increased geothermal fluid production by boosting the bulk permeability and fluid storage of a production zone. However, the hydromechanical properties of faults are inherently heterogeneous and anisotropic, making it challenging to distinguish between permeable and impermeable faults prior to drilling and testing. The probability of locating zones of enhanced fault permeability can be improved by preliminary fault stress state modeling. Such modeling can reduce the uncertainty and risk of exploring for fault-related geothermal targets, particularly in early stage exploration projects or in areas of unknown or complex geology [46].

One recent development has been the realization by the conventional geothermal industry that the heat

Mitigation potential of geothermal power plants in the world based on data of Table 6 and assumptions for emission of 120 g CO₂/kWh for today and 10 g CO₂/kWh for future technology (EGS).



Engineered Geothermal Systems, Development and Sustainability of. Figure 4

CO₂ Emission from various energy resources

available in the vicinity of hydrothermal systems is significantly greater than what is produced from aquifers within those fields. The knowledge acquired from the development of EGS is being applied to extend existing hydrothermal fields to access much larger volumes of hot rock mass to contribute to the power production. Approaches include stimulations of existing dry wells and drilling new wells to connect to the hydrothermal reservoir, provided the wells are located in a favorable direction with respect to the local stress regime. ORMAT Technologies is carrying out pioneering work in this field at Desert Peak near Reno, USA. There is tremendous potential for the take-up of this technology once the concept is demonstrated.

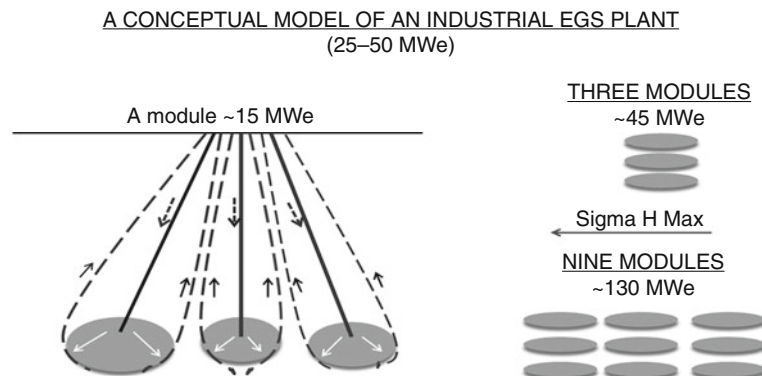
The present economic conditions in many parts of the world mean that a combination of heat and power is likely to generate higher revenue for EGS plants rather than just power alone. This has been the policy pursued by some EGS companies, e.g., EGS Energy Ltd (www.egs-energy.com). The use and distribution of heat may not be applicable everywhere, but a serious consideration is essential to increase revenue and economic sustainability of projects.

Around 3–6 MWe EGS power plants are a starting point in the initial take-up of the technology, but much larger power generating capacity (25–50 MWe) must be demonstrated before the technology can be considered mature and acceptable as a big power player. In order to achieve such a goal, a concept put forward by EGS Energy Ltd, is to successfully replicate EGS reservoirs again and again. Taking the example of the achievement at Soultz and assuming that the

connection between GPK3 and GPK4 could have had similar characteristics to that between GPK2 and GPK3, the module of one injector and two producers can be replicated again and again in parallel and serial configuration. Each module could produce around 5 MWe, so three stacked in series could give ~15 MWe (Fig. 5). Three 15 MWe systems stacked in parallel would produce ~45 MWe. This configuration could be extended/developed to generate a larger power the range of up to 100 MWe or more. An important criterion in developing such strategy is the uniformity of stress and thermal gradient so that once a prototype with appropriate characteristics has been developed and tested, it could be replicated with relative confidence. Minor occasional variations in characteristics can be rectified using an appropriate tool kit of responses as necessary.

Sustainability of EGS Systems

Sustainability can be briefly defined as “the capacity to endure, hold up or maintain.” In the current atmosphere of community concern for the environment, the question of sustainability arises for any technology that is being developed for the future. For EGS, sustainability could relate to maintaining the temperature and pressure in the reservoir, the electricity output from the power plant, the economic return on investment, the lifespan of surface plant, or other measures. At present, the question of long-term sustainability for EGS is virtually impossible to quantify. Only one EGS commercial system is in existence, and this has been



Engineered Geothermal Systems, Development and Sustainability of. Figure 5
Scaling up scenarios for producing 15–130 MWe (from EGS Energy Ltd)

operating only in the last few years. The question of sustainability still arises, however, and one can possibly address this at present through preliminary observations [47], numerical modeling, and value judgment from working in the development of such systems for over 30 years.

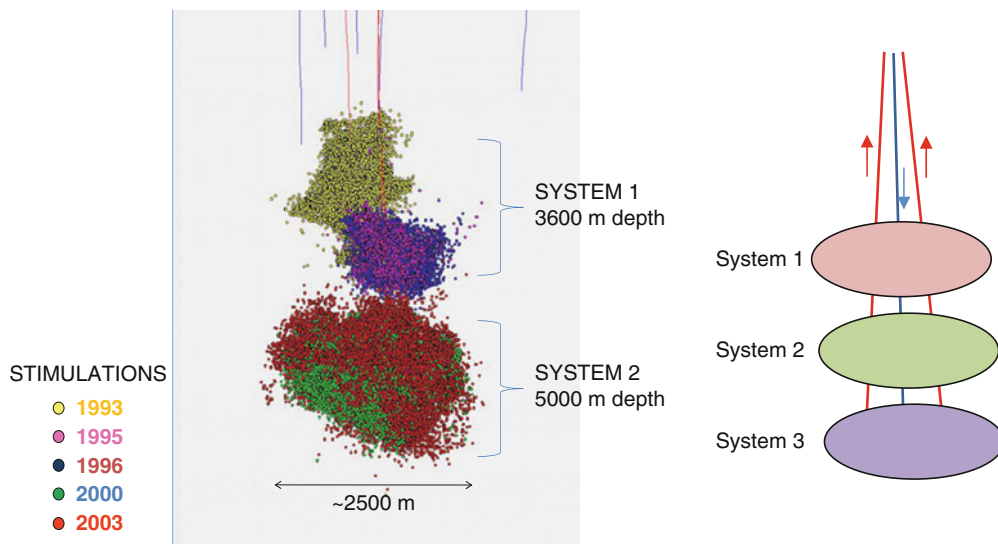
In an EGS power plant, one assumes the term “endurance” is aimed at the available thermal energy in the reservoir being able to recharge itself in a sufficient time to be used again in the future, and not associated with the mechanical items such as the casing of the well or the power plant. Sustainability questions will be raised when attempts are made to improve or extend the life of an existing EGS reservoir to conserve resource and/or improve profitability.

For example, calculations suggest that the recovery of the depleted heat in a fractured rock reservoir may take around 75 years to recover by the natural conduction process from the surrounding rocks [12]. This assumes that no convection process is involved, and all the heat transfer is by conduction. In order to extract heat from the upper crust using EGS methods, existing fractures must be opened up by a shearing mechanism to improve fluid flow. In this case, the fractured volume may internally act as convective cells [48]. A convection

process is more efficient in transferring heat, and if this is the case, then it is possible that a depleted EGS reservoir might thermal recover in a shorter time than by conduction alone.

One of the observations during the 5-year circulation at the Rosemanowes site was that as the center part of the reservoir was cooling [49], causing a shrinking of the rock mass, the outer part of the reservoir was being opened up and played a greater part in the extraction of heat by the creation of new flow paths. This was concluded from tracer tests that suggested the generation of new flow paths. This suggests that if one manages the reservoir correctly, then there is a good possibility that additional channels could open up through time and thus improve sustainability.

Experience at Soultz demonstrates that a number of reservoirs can be created within the same granite body, above or beside each other (Fig. 6), depending on the stress regime. The upper reservoir (3,600 m) at Soultz was circulated initially for a while and was then sealed off. A new reservoir was created below the first at 5,000 m depth. This same concept could be used to create not just two but three EGS reservoirs, one stacked above the other. The bottom reservoir (system 3) would be exploited first. When it had been sufficiently depleted,



Engineered Geothermal Systems, Development and Sustainability of. Figure 6
Stacking up of EGS reservoirs (an example from Soultz and a conceptual model)

the reservoir would be isolated and the one above it opened and used for generating power. Once the second one is depleted, then that would also be isolated and the upper reservoir would be used. Assuming that each reservoir lasted for around 25 years, by the time the upper reservoir is depleted, the bottom reservoir would be recharged to almost the original temperature through conduction from the surrounding rocks. This suggests that the wells have to be completed in such a way that a number of EGS reservoirs can be created one above the other, such that any one can be isolated to operate independently of the others. A mechanism like this may make it possible for EGS reservoirs to be sustainable for a century or more.

Although EGS projects can offer a viable alternative renewable energy source, they should not be considered in isolation from other local constraints. For example, the creation and operation of an EGS project using a water-cooled power plant requires a local reliable source of water at reasonable cost. Furthermore, EGS field projects carried out throughout the world have experienced the impact of water losses through trial production testing. Controlling water losses is an important aspect of sustainability for a geothermal energy project particularly in areas of restricted water availability such as Geodynamic's Habanero project located in arid outback Australia. Water losses can have significantly negative economic and environmental impacts if not managed. Techniques proposed to control water losses include minimizing injection pressures to control reservoir growth during circulation, the use of injection proppants, and improving hydraulic connectivity between wells through the use of techniques such as "focused stimulation" [12, 50].

In hydrothermal systems, operators need to manage the system not just to prolong its life but also to make it economically attractive. This requires characterization of the reservoir using tracers to monitor the change in the reservoir properties and evaluation of the geochemistry to manage dissolution and precipitation effects. Geochemistry aspects associated with EGS reservoirs are still uncertain and not yet quantified [51, 52]. This is again a function of there being no long-term operating plants yet. Some of the experiments done at the Soultz site during various circulation tests suggest that dissolution, precipitation, and corrosion may be

controlled by controlling the pH of the circulating fluid, maintaining back pressure on the production flow, introducing anti-corrosion agents in the system, and by pumping the production fluid back in to the formation without exposing the fluid to the open environment.

There has been some concern in the past that preferential flow paths may develop as the reservoir cools, and that this may cause the abandonment of a reservoir earlier than planned. This to some degree has been addressed by increasing the distances between injector and producer wells from ~200 m to over 600 m. The experience at Soultz suggests that these distances may be further increased to over 1,000 m if one uses a focused stimulation method. Focused stimulation allows the pressure to build up rapidly between the wells to shear existing joints at distances some hundreds of meters from any one well, and also takes significantly less time and volume of water. The result is a more consistent system of stimulated joints throughout the rock volume, and hence a lower probability of short circuits. Additionally, tracer tests carried out regularly in the reservoir would give an early indication of the development of preferential flow paths (short circuits), and then methods could be employed to seal them using various chemicals. Such sealing agents (calcite based) are currently being developed to seal a path temporarily or permanently. This, in principle, should help to extend the life and sustainability of a reservoir.

As mentioned earlier, sustainability is an issue that cannot completely be quantified until there are a number of operating EGS reservoirs for a long time, but the indications are that there is every probability that most issues affecting sustainability can be managed to provide an optimal return on investment.

Possible Road Map for EGS

There is a tremendous quantity of heat stored in the Earth ($\sim 10^{13}$ EJ, [47]), predominantly from the decay of natural radioisotopes. Some of this heat resource is being successfully exploited all over the world at the tectonic boundaries where the crust is thin. These resources have, to a great extent, large volumes of very hot fluid trapped underground and are accessed

by drilling in to them (Iceland, New Zealand, etc.). These resources are classified as a hydrothermal resource, but there are extensive areas of the world where the crust is not so thin, but extensive heat energy is available at greater depth.

EGS technology provides a mechanism to exploit this available heat resource at great depth by drilling in to it and creating an artificial reservoir by pumping fluid under pressure, opening up natural fractures to allow cold fluid to flow through it, and extract the heat to drive turbines at the surface. Only recently the technological and financial support was available to create such a system with relevant parameters to make it commercially viable. This was achieved under an international research environment which facilitated funds and scientists to address these problems and solve them.

International Energy Agency has recognized the potential of EGS worldwide and the cutting edge geothermal research required to make it successful. IEA has therefore set up a team to formulate a road map to assist the exploitation of this technology.

I suspect some of the aspects covered for the road map will be:

- What geophysical technique will help in selecting an ideal EGS site?
- How can one determine stress values and gradient in that specific area?
- How can one encourage joint international research to address some of the uncertainties, such as the behavior of an EGS reservoir in different stress regimes; how long will the reservoir last, etc.
- Can one support new drilling techniques to reduce the cost of drilling? Drilling can be over half of the cost of development of an EGS Project
- Facilitate a procedure of fast tracking for planning permission to reduce effective cost.
- Understand or develop mitigating methods to reduce large induced seismic events in EGS plants.
- Provide a mechanism such as insurance scheme to assist in any large drilling accidents.
- Long-term international research on the behavior of EGS reservoir and the control of likely geochemical, dissolution, and precipitation problems.
- Improve binary power plants to improve power conversion efficiency and reliability.

Future Directions

EGS has gone through the fundamental research to scientific demonstration stage, and it is now ready for application in commercial environment, but, it does need supporting financial incentive to encourage potential investors.

These can be:

- A feed in tariff similar or better than that set by Germany will encourage potential investors to be prepared to take some risk.
- Additionally, the potential of being able to provide an abundant supply of hot water from an EGS plant has not been taken seriously, particularly as it can provide this 24 h a day and almost 95% of the time in a year.
- Environment friendly premium, as it has virtually no CO₂ emissions or any major pollutants.
- A separate feed in tariff for the supply of hot water.
- EGS has a potential for becoming a provider of large-scale power (200MWe +) and heat. This can become a strategic power base and additional financial support from government will help investors to expand in this direction.
- Insurance schemes to protect against unforeseen accidents during drilling.
- Establishment of fast tracking method in obtaining planning and fulfilling local regulations. This can take a long time and can become a major burden to investors at an early stage of development.

Acknowledgements

The authors would like to thank all the funding agencies that helped to establish and support the Soultz project, including the European Commission, Agence de l'Environnement et de la Maîtrise de l'Energie (France), the German Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit within the frame of the "Zukunftsinvestitionsprogramm," the Projektträger of the Forschungszentrum Jülich (Germany), the Department of Trade and Industry (DTI) (UK), and the Members of the EEIG "Exploitation Minière de la Chaleur."

The authors would also like to thank all the teams who contributed to the success of the projects at Rosemanowes and Soultz for over 30 years of research.

In particular, the authors would like to give special acknowledgement to Perry Moore (SII) who died on March 21, 2003, in Kazakhstan, and who played a significant part in the success of the Rosemanowes and Soultz projects.

Bibliography

- Baria R, Baumgaertner J, Gérard A, Kappelmeyer O (1992) HDR project at Soultz-sous-Forêts. *Geotherm Res Counc Trans* 18:387–394
- Abe H, Niitsuma H, Baria R (1999) Hot dry rock/Hot wet rock academic review: special issue. *Geothermics* 28(4/5). Pergamon Press, ISSN 0375-6505
- Garnish JD (1985) Hot dry rock – a European perspective. In: *Proceedings geothermal research council's international symposium on geothermal energy, international volume, Hawaii*, pp 329–337
- Baria R, Baumgaertner J, Gerard A, Jung R, Garnish J (1999) European HDR research program at Soultz-sous-Forêts (France) 1987–1996. In: Abe H, Niitsuma H, Baria R (eds) *Geothermics special issue, Hot dry rock/Hot wet rock, Academic review*, 28(4/5). pp 665–669, ISSN 0375-6506
- Baria R, Baumgaertner J, Gérard A, Jung R, Garnish J (1998) The European HDR research program at Soultz-Sous-Forêts (France). In: Baria R, Baumgaertner J, Gerard A, Jung R (ed) *International conference 4th HDR forum, Proceeding, 2002, Strasbourg, France*, ISBN 3-510-95859-4
- Hot Dry Rocks (2011) Summary of process and data sources used in the development of the EGS Potential maps and tables of Australia. Unpublished report (available from www.google.org/egs), 20 p
- Smith MC (1975) The potential for the production of power from geothermal resources – Los Alamos scientific laboratory report, LA-UR-73-926
- Barton CA, Zoback MD, Moos D (1995) Fluid flow along potentially active faults in crystalline rock. *Geology* 23:683–686
- Finkbeiner T, Barton CA, Zoback MD (1997) Relationships among in-situ stress, fractures and faults, and fluid flow: Monterey formation, Santa Maria Basin. *AAPG Bull* 81(12):1975–1999
- Barton CA, Hickman S, Morin R, Zoback MD, Finkbeiner T, Sass J, Benoit D (1997) Fracture permeability and its relationship to in-situ stress in the Dixie Valley, Nevada, geothermal reservoir. In: *Proceedings 22nd workshop on geothermal reservoir engineering Stanford University, Stanford*, 27–29 Jan 1997
- Bruel D (1997) Heat exchange modelling using a discrete fracture network model with thermo-mechanical interactions – evaluation of the thermal performance of the reservoir developed at Soultz-sous-Forêts in 1996, ARMINES
- M.I.T. Report (2006) The future of geothermal energy, impact of enhanced geothermal systems (EGS) in the United States in 21st century. Massachusetts institute of technology report. ISBN: 0-615-13438-6. (<http://geothermal.inel.gov>)
- Beardmore GR, Rybach L, Blackwell D, Baron C (2010) A protocol for estimating and mapping global EGS potential. *GRC Trans* 34:301–312
- Potter RM, Robinson ES, Smith MC (1974) Methods of extracting heat from dry geothermal reservoirs, US Patent #3,786858
- Brown DW, Duchane DV (1999) Scientific progress on the Fenton hill HDR project since 1983. In: Abe H, Niitsuma H, Baria R (eds) *Geothermics special issue, Hot dry rock/Hot wet rock, Academic review*, 28(4/5), pp 591–601, ISSN 0375-6506
- Garnish JD (1976) Geothermal energy: the case for research in the United Kingdom. Energy paper no. 9, London, HMSO, 66 pp
- Pine RJ, Batchelor AS (1984) Downward migration of shearing in jointed rock during hydraulic injections. *Int J Rock Mech Min Sci* 21(5):249–263
- Baria R, Hearn KC, Batchelor AS (1985) Induced seismicity during the hydraulic stimulation of the potential hot dry rock geothermal reservoir – submitted to the fourth conference on acoustic emission/microseismic activity in geology structures and materials, Pennsylvania State University, 22–24 Oct 1985, 26 pp
- Baria R, Green ASP (1989) Microseismics: a key to understanding reservoir growth. In: Roy Baria (ed) *Hot dry rock geothermal energy. Proceedings Camborne school of mines international hot dry rock conference, Camborne School of Mines Redruth, UK*, 27–30 June 1989. Robertson Scientific Publications, London, 1990, pp 363–377, ISBN 1-85365-217-2
- Parker RH (1989) Hot dry rock geothermal energy. Phase 2B final report of the Camborne School of Mines project, vol 1 & 2. Pergamon, Oxford. ISBN 0-08-037929-X
- Batchelor AS (1989) Hot dry rocks and its relationship to existing geothermal systems. In: Baria R (ed) *Hot dry rock geothermal energy proceedings Camborne school of mines international hot dry rock conference, Camborne School of Mines Redruth*, 27–30 June 1989. Robertson Scientific Publications, London, 1990, pp 13–29, ISBN 1-85365-217-2
- Garnish JD (1986) Introduction: background to the workshop. In: Garnish J (ed) *Proceedings of the first EEC/US workshop on geothermal hot dry rock technology, geothermics special issue*, 16(4), 1987. Brussels, pp 323–330
- Gérard A, Kappelmeyer O (1987) The Soultz-sous-Forêts Project. In: Garnish J (ed) *Proceedings of the first EEC/US workshop on geothermal hot dry rock technology, geothermics special issue*, 16(4), 1987, pp 393–399
- Baumgärtner J, Gérard A, Baria R, Jung R, Tran-Viet T, Gandy T, Aquilina L, Garnish J (1998) Circulating the HDR reservoir at Soultz: maintaining production and injection flow in complete balance. In: *Proceedings 23rd workshop of geothermal reservoir engineering, Stanford University, Stanford*
- Baria R, Baumgärtner J, Gérard A (1993) Heat mining in the Rhinegraben. Socomine internal project report
- Baria R, Garnish J, Baumgärtner J, Gérard A, Jung R (1995) Recent development in the European HDR research programme at Soultz-sous-Forêts (France). *Proceeding of the*

- world geothermal congress, Florence, International Geothermal Association, vol 4, pp 2631–2637, ISBN 0-473-03123-X
27. Baumgärtner J, Jung R, Gérard A, Baria R, Garnish J (1996) The European HDR project at Soultz-sous-Forêts: stimulation of the second deep well and first circulation experiments. In: Proceedings 21st workshop of geothermal reservoir engineering, Stanford University, Stanford, California, SGP-TR-151, pp 267–274
 28. Gérard A, Baumgärtner J, Baria R (1997) An attempt towards a conceptual model derived from 1993–1996 hydraulic operations at Soultz. In: Proceedings of NEDO international geothermal symposium, Sendai, vol 2, pp 329–341
 29. Shock RA (1986) An economic assessment of hot dry rocks as an energy source for the UK – Energy Technology Support Unit Report ETSU-R-34. Publ. HMSO (London), 215 p
 30. Geodynamics (2010) Geodynamics company update. In: Proceedings Australian geothermal conference 2010, Adelaide, Australia, 16–19 Nov 2010
 31. Heidinger P, Dornstädter J, Fabritius A (2006) HDR economic modelling: HDRec software. *Geotherm* 35:683–710
 32. Heidinger P (2010) Integral modeling and financial impact of the geothermal situation and power plant at Soultz-sous-Forêts. *Comptes Rendus Geosci* 342: 626–635. Elsevier Masson SAS
 33. Abé H, Mura T, Keer LM (1976) Growth rate of a penny-shaped crack in hydraulic fracturing of rocks. *J Geophys Res* 81(29):5335–5340
 34. Duffield RB, Nunz GJ, Smith MC, Wilson MG (1981) Hot dry rock – geothermal energy development program, annual report FY80, Los Alamos National Laboratory report, LA-8855-HDR, July 1981, 211 p
 35. Hubbert MK, Willis DG (1957) Mechanics of hydraulic fracturing. *Pet Trans AIME* 210:153–168
 36. Genter A, Fritsch D, Cuenot N, Baumgärtner J, Graff J-J (2009) Overview of the current activities of the European EGS Soultz project: from exploration to electricity generation. In: Proceedings thirty-fourth workshop on geothermal reservoir engineering, Stanford University, Stanford, 9–11 Feb 2009
 37. Baria R, Michelet S, Baumgaertner J, Dyer B, Gerard A, Nicholls J, Hettkamp T, Teza D, Soma N, Asanuma H, Garnish J, Megel T (2004) Microseismic monitoring of the world's largest potential HDR reservoir. Twenty-ninth workshop on geothermal reservoir engineering, Stanford University, Stanford, California, 26–28 Jan 2004
 38. Baria R, Michelet S, Baumgärtner J, Dyer B, Nicholls J, Hettkamp T, Teza D, Soma N, Asanuma H, Garnish J (2005) Creation and mapping of 5000m deep HDR/HFR reservoir to produce electricity. In: Proceedings WGC 2005 (CDROM), Antalya
 39. Baria R, Jung R, Tischner T, Nicholls J, Michelet M, Sanjuan B, Soma N, Asanuma H, Dyer B, Garnish J (2006) Creation of an HDR reservoir at 5000m depth at the European HDR project. Thirty-first workshop on geothermal reservoir engineering. Stanford University, Stanford, 30 Jan–1 Feb 2006
 40. Hettkamp T, Baumgaertner J, Baria R, Gerard A, Gandy T, Michelet S, Teza D (2004) Electricity production from hot rocks. In: Proceedings 29th workshop of geothermal reservoir engineering, Stanford University, Stanford, 26–28 Jan 2004, pp 184–193
 41. Brown DW, DuTeaux R, Kruger P, Swenson D, Yamaguchi T (1999) Fluid circulation and heat extraction from engineered geothermal reservoirs. In: Abe H, Niitsuma H, Baria R (eds) *Geothermics special issue, hot dry rock/hot wet rock, Academic review*, 28(4/5), pp 553–572, ISSN 0375-6506
 42. Sanjuan B, Pinault, Rose P, Gerard A, Brach M, Braibant G, Crouzet C, Foucher J-C, Gautier A, Touzelet S (2006) Tracer testing of the geothermal heat exchanger at Soultz-sous-Forêts (France) between 2000 and 2005. *Geotherm* 35 (5–6):622–653
 43. Majer E, Baria R, Stark M (2008) Protocol for induced seismicity associated with enhanced geothermal systems. Report produced in Task D Annex I (9 April 2008), International Energy Agency-Geothermal Implementing Agreement. <http://www.iea-gia.org/publications.asp>
 44. Major E, Baria R, Stark M, Oates S, Bommer J, Smith B, Asanuma H (2007) Induced seismicity associated with Enhanced Geothermal Systems. *Geotherm* 36(2007):185–222
 45. Helm JA, Hoang-Trong P (1993) Assessing the natural and induced seismic risk associated with a hot dry rock (HDR) geothermal energy project with specific reference to Soultz-sous-Forêts, N.E. France. American Geophysical Union, 1993 fall meeting, San Francisco
 46. Mortimer L, Adnan, A, Simmons CT (2010) Targeting faults for geothermal fluid production: exploring for zones of enhanced permeability. In: Proceedings Australian geothermal conference 2010, Adelaide, 16–19 Nov 2010
 47. Rybach L, Mongillo M (2006) Geothermal sustainability—a review with identified research needs. *GRC Trans* 30:1083–1090
 48. Bächler D, Kohl T, Rybach L (2003) Impact of Graben-parallel faults on hydrothermal convection – Rhine Graben case study. *Phys Chem Earth* 28:431–441
 49. Parker RH (1989) Geothermal energy at the Camborne school of mines. *Geotherm Counc Bull* 18(9):3–7
 50. Cordon E, Driscoll JP (2008) Full life-cycle water requirements for deep geothermal energy developments in South Australia. Unpublished report to the Department of Primary Industries and Resources, South Australia. 50p. http://www.pir.sa.gov.au/_data/assets/pdf_file/0018/110556/TIG_4_PIRSA_Water_Project_26May09.pdf
 51. André L, Rabemanana V, Vuataz F-D (2006) Influence of water-rock interactions on fracture permeability of the deep reservoir at Soultz-sous-Forêts, France. *Geotherm* 35(5–6):507–531
 52. Vuataz F-D, Durst P (2001) Fluid-rock interactions and geochemical modelling of the formation brine in the fissured reservoir of the Soultz-sous-Forêts HDR test site. Final report, Federal office for education and science, project OFES N° 98.0008-3

Environmental Countermeasures and Restoration

SERGEY V. FESENKO¹, BRENDA J. HOWARD²

¹Department of Nuclear Sciences and Applications, IAEA Environment Laboratories, International Atomic Energy Agency, Vienna, Austria

²Centre for Ecology and Hydrology, Lancaster Environment Centre, Bailrigg, Lancaster, UK

Article Outline

Glossary (According to the IAEA nuclear safety glossary [1])

Definition of the Subject and Its Importance

Introduction

Origin and Characteristics of Contaminants in the Environment

International Frameworks for Application of Countermeasures and Remediation of Contaminated Areas

Site Evaluation for Remediation Purposes

Management Options for the Application of Countermeasures and Remediation for Different Environments

Countermeasures and Remediation in Areas Affected by Radiation Accidents

Remediation of Sites Affected by Nuclear Testing

Remediation of Sites within Decommissioning of Facilities

Decision-Aiding Tools

Future Directions

Bibliography

Glossary (According to the IAEA nuclear safety glossary [1])

Acceptable limit A limit acceptable to the relevant regulatory body.

Accident Any unintended event, including operating errors, equipment failures or other mishaps, the consequences or potential consequences of which are not negligible from the point of view of protection or safety.

Action level The level of dose rate or activity concentration above which remedial actions or protective

actions should be carried out in chronic exposure or emergency exposure situations.

Contamination Radioactive substances on surfaces, or within solids, liquids, or gases (including the human body), where their presence is unintended or undesirable, or the process giving rise to their presence in such places.

Cost-benefit analysis A systematic economic evaluation of the positive effects (benefits) and negative effects (disbenefits, including monetary costs) of undertaking an action.

Countermeasure An action aimed at alleviating the radiological consequences of an accident.

Decommissioning Administrative and technical actions taken to allow the removal of some or all of the *regulatory controls* from a *facility*.

Decontamination The complete or partial removal of contamination by a deliberate physical, chemical, or biological process.

Dose A measure of the energy deposited by radiation in a target.

Emergency A nonroutine situation or *event* that necessitates prompt action, primarily to mitigate a hazard or adverse consequences for human health and safety, quality of life, property, or the environment.

Evacuation The rapid, temporary removal of people from an area to avoid or reduce short-term radiation exposure in an emergency.

Exemption level A value established by a regulatory body and expressed in terms of activity concentration, total activity, dose rate, or radiation energy, at or below which a source of radiation may be granted exemption from regulatory control without further consideration.

Exposure pathway A route by which radiation or radionuclides can reach humans and cause exposure.

Exposure The act or condition of being subject to irradiation.

Intervention Any action intended to reduce or avert exposure or the likelihood of exposure to sources that are not part of a controlled practice or that are out of control as a consequence of an accident.

Justification (for intervention) The process of determining whether a proposed *intervention* is likely,

overall, to be beneficial, as required by ICRP's *System of Radiological Protection*, i.e., whether the benefits to individuals and to society (including the reduction in *radiation detriment*) from introducing or continuing the *intervention* outweigh the cost of the *intervention* and any harm or damage caused by the *intervention*.

Protective action An intervention intended to avoid or reduce doses to members of the public in emergencies or situations of chronic exposure.

Reference level An action level, intervention level, investigation level, or a recording level.

Relocation The nonurgent removal or extended exclusion of people from a contaminated area to avoid chronic exposure.

Remedial action Action taken when a specified action level is exceeded, to reduce radiation doses that might otherwise be received, in an intervention situation involving chronic exposure.

Remediation Any measures that is carried out to reduce the radiation exposure from existing contamination of land areas through actions applied to the contamination itself (the source) or to the exposure pathways to humans.

Risk A multi-attribute quantity expressing hazard, danger, or chance of harmful or injurious consequences associated with actual or *potential exposures*. It relates to quantities such as the probability that specific deleterious consequences may arise and the magnitude and character of such consequences.

Risk assessment Assessment of the radiological risks associated with normal operation and possible accidents involving a source or practice.

Site evaluation Analysis of those factors at a site that could affect the safety of a facility or activity on that site.

Waste management All administrative and operational activities involved in the handling, pretreatment, treatment, conditioning, transport, storage, and disposal of radioactive waste.

Definition of the Subject and Its Importance

There are a large number of sites around the world contaminated with radionuclides which, to varying extents, require attention. Some of these sites were contaminated by nuclear testing or radiation accidents while

others became contaminated as a result of authorized discharges (e.g., hospitals) or inadequate waste management and/or waste disposal practices. In recent years, problems associated with radioactively contaminated sites have led to considerable concern amongst both the general public and decision makers. To a large extent, the concern has been stimulated by the occurrence of several radiation accidents, most notably the Chernobyl accident. These issues have also gained a particular importance because of an increased public concern about the safety of radioactive waste disposal and the associated justification of a sustainable development of nuclear power.

The first stage of any strategy responding to contamination is an evaluation of the need for action. Decisions to use any actions to protect the population or remediate the environment are normally based on a comprehensive site evaluation and justification of the intervention, if necessary. This analysis should then be able to identify appropriate management options.

Many factors govern the effectiveness and cost of countermeasures or remedial actions and need to be evaluated to identify a suitable strategy. The contributions of these factors are site specific, varying with the characteristics of the contamination and the receiving environment as well as on factors such as local and national priorities for remediation planning.

Environmental restoration is often considered as a system of actions having a final goal to return a site to the conditions that prevailed before the contamination. However, in practice, this is often not feasible, especially if vast areas are affected and require a response.

Planning of environmental countermeasures or remedial actions is a complicated, multidimensional task which needs to take account of radiological, ecological, and socioeconomic issues within a holistic approach. The challenge is to identify suitable, effective, individual actions or procedures (such as remediation technologies, communication strategies, or decision support technologies) within a holistic consideration of the entire range of available countermeasures and remediation actions.

Introduction

Contamination of the environment to an extent that requires countermeasures or remediation may occur

due to many different reasons. Obvious sources of radioactive contamination which may require the use of countermeasures include radiation accidents or incidents, but remediation may also be needed for radiation legacy sites [2–7]. Radioactive contamination can be caused inadvertently by human activities involving processes in which natural radionuclides can become concentrated in areas not normally controlled by regulatory bodies. Such activities can include conventional mining and processing of ores, production of phosphogypsum, etc. [8]. There are also many other sources of radioactive contamination such as the application of radionuclides in industry, research, and medicine. The responses to various accidents and existence of some highly contaminated legacy sites have required the development of a variety of countermeasure and remediation management options to protect human health or to remediate the environment.

According to the current international terminology, contamination refers to radioactive substances which occur on surfaces, or within solids, liquids, or gases (including the human body), where their presence is unintended or undesirable [1]. However, almost all human activities are accompanied by the release of contaminants into the environment and such releases are normally under strict regulatory frameworks.

The terms *countermeasure* and *remediation* are frequently used, but often applied differently by various authors. Such a situation can often lead to misinterpretations in descriptions of the activities applied to decrease radiation impact on the public and the environment in an emergency or in exposure situations when contamination is already present. To clarify the issue, the terms here are used consistently in the following way. Management options which have been applied in the case of emergency are countermeasures, and they can include both protective and remedial actions [1]. However, the same management options when applied in the situation of existing exposure are considered remedial actions. Existing exposure in this situation includes options applied when an emergency exposure situation has been declared to have ended or in areas contaminated by residual radioactive material from past practices, and are part of a *remediation process*. The more informal term *environmental countermeasure* is also used elsewhere for management options used as remedial actions.

There is no definitive set time period when an option will change from a countermeasure to a remediation in the case of emergencies. The transition from an emergency exposure situation to a subsequent existing exposure is characterized by a change in management, from countermeasure strategies mainly driven by urgency, with potentially high levels of exposures and predominantly central decisions to more decentralized remediation strategies aiming to improve living conditions and reduce exposures as low as reasonably achievable given the circumstances [9]. These strategies must take into account the long-term dimension of the situation with the possible direct involvement of the exposed individuals in their own protection [10].

There are some management option countermeasures such as evacuation or application of iodine tablets, which can be considered only as protection actions and never as remedial actions since they are only used in the acute or early phase of an emergency. Conversely, some management options can be used both as countermeasures and remedial actions, depending on the situation or interpretation of a purpose of the application, such as clean feeding of animals to decontaminate them.

A wide range of countermeasures and remedial actions have been developed, tested, and implemented in contaminated areas. As a result, a large amount of data on their effectiveness has been generated, together with information on ancillary factors such as the required resources and costs. These measures vary considerably in their effectiveness, cost, and practicability and this difference has led to a need to develop a variety of approaches and adequate decision-aiding technologies and Environmental Decision Support Systems (EDSSs) that can be effectively used for countermeasure or remediation planning and optimized application of such actions [11–13].

The main sources of the data on the effectiveness of management options has been the application of countermeasures in areas affected by major radiation accidents, remediation of nuclear testing sites, and areas affected by nuclear facilities or other human nuclear activities such as the use of radionuclides for medical, research, or industrial purposes. Analysis of the lessons learned from these case studies has been invaluable in demonstrating that the efficiency of protection of management options depends on many factors, including environmental conditions, properties of the radionuclides, management on the affected areas, and

response from the local population and stakeholders [11, 14–17]. This leads to a need to adopt a holistic approach, the application of which is to be facilitated by modern regulatory frameworks for countermeasures and remediation of contaminated areas [10], especially with a target of sustainable development of the environment.

Origin and Characteristics of Contaminants in the Environment

There are many different sources of radioactive substances in the environment. At present, the most important sources which have resulted (or which still result) in the need for the application of countermeasures or remediation can be divided into two main groups. The first group can be termed *radiation legacy*: this group includes military test explosions, radiation accidents, nuclear weapon accidents, early design nuclear fuel cycle facilities, and nuclear applications with inadequate operational and disposal practices [2, 4]. In some of these cases, the application of countermeasures was necessary to protect the population of the affected areas before remediation of these sites was considered.

The second group consists of *existing* human activities including uranium mining, processing of ore, fuel fabrication, energy generation, reprocessing of nuclear fuel, and various other applications of radionuclides in industry, research, and medicine. For all these activities countermeasures are not required because the corresponding discharges of radionuclides into the environment or the measured concentrations of radionuclides in the environment are below the acceptable limits. However, remediation is a planned and necessary part of the decommissioning of these facilities and the sites should be restored to the status which (under existing regulations) allows release of the site to unrestricted use (ideally) or for some specific purposes (restricted use) if the first option is not feasible.

Radiation Legacy Sites

Most sites requiring large-scale remediation are part of a “radiation legacy” [4]. Radionuclides of concern at such sites have a long physical half-life, high radiological toxicity, and high mobility in the environment or are elements (including analogs) which have an essential role in the biosphere [18].

During the period 1940–1970, many sites were heavily contaminated because of the prevailing radioactive waste disposal practices at facilities for nuclear weapons material production. In particular, large quantities of radioactive liquid wastes, sometimes mixed with other chemicals, were commonly disposed of directly into the ground, surface water bodies, or marine systems (for instance at the Mayak, Hanford, and Sellafield production facilities). These disposal practices led, in some instances, to serious environmental problems which required, or still require, remediation [2, 4, 17].

Another major legacy group arose from nuclear explosives, which have been detonated above ground, on the surface, and below ground at many test sites around the world. Nuclear weapons accidents have occurred around the world since the 1950s. Severe accidents include the losses of nuclear weapons from aircraft at Palomares (Spain) in 1966 and Thule in Greenland in 1968 [17, 19–21].

Additionally, about 100 civilian peaceful nuclear explosives have been performed in the former Soviet Union countries and in the USA for different purposes (constructions of dam, gas reservoirs, etc.). Although the contribution of these explosives to the global dose was very low [5], in some local areas contamination of the environment has required application of remedial actions [6, 22, 23].

Severe radiation accidents at the Mayak radiochemical plant (Kyshtym accident) in 1957 and the Chernobyl nuclear power plant in 1986 have led to significant contamination by long-lived radionuclides over many years, requiring long-term remediation [17, 24–26] after the acute phase.

In some cases, either through accidents or a lack of radiation safety culture, sites have been left contaminated with residues even from authorized operations. Such sites include factories where radium was used in luminescent paint and thorium was used in thorium-coated gas mantles [4].

Existing Nuclear Activities and Current Sources of Radioactive Contamination

Due to an improved radiation safety culture over the last several decades, the discharge of radionuclides into the environment has greatly declined in many countries [5]. Despite low discharges of radionuclides, nuclear

fuel cycle facilities still contribute to contamination of the environment (mainly within the facility) to a sufficient extent that remedial actions might be advisable even in the absence of radiation accidents. Naturally occurring radioactive materials (NORMs) such as ^{238}U , ^{232}Th , and their progeny are often the most significant environmental contaminants at such sites. NORMs can also occur in many ores containing metals such as copper, gold, coal, and monazite. Residues from the production of phosphoric acid and various phosphates may also cause contamination of the environment [27]. The combustion of coal for electricity generation, smelting processes, and heating leads to the release and dispersion of some natural radionuclides, mainly radium, in the environment.

Normally, remediation of all such sites is an obligatory part of the decommissioning process and is subject to special compliance monitoring [17, 28] which must prove that affected areas have been adequately remediated before it is permissible for them to be used for other purposes.

Currently, radionuclides are extensively used for a variety of scientific, medical, agricultural, and industrial purposes [2]. The major users of unsealed radioactive sources are hospitals where nuclear practices are used for diagnostic and therapeutic purposes, and to a lesser extent for treatments. Some of the radionuclides used for such activities (^3H , ^{14}C , ^{32}P , ^{35}S , and ^{131}I) occur in the liquid effluents which can be discharged into the environment [6]. These radionuclides are also in use in a range of research applications such as genetic studies, animal health or environmental studies, while ^{147}Pm is widely used in industry [4].

An additional long-term source of contamination of the environment is radioactive waste. Environmental contamination may occur during the collection, processing, storage, and disposal of wastes associated with the use of radioactive isotopes for industrial, research, and medical applications [4, 6].

International Frameworks for Application of Countermeasures and Remediation of Contaminated Areas

The major source of international recommendation on this subject is the International Commission on Radiological Protection (ICRP). The ICRP is the primary

international body that recommends the fundamental principles and quantitative basis upon which appropriate radiation protection measures can be established. It leaves the responsibility of formulating the specific advice, codes of practice, or regulations tailored to the individual countries (<http://www.icrp.org/about.asp>).

Overall, the ICRP's recommendations serve as a major basis for all protective actions, including environmental countermeasures and any remedial actions. International standards and supporting technical documents, such as the International Basic Safety Standards for Protection Against Ionizing Radiation and for the Safety of Radiation Sources [29], are also based on ICRP recommendations.

The ICRP recently revised its approach to characterizing situations in which exposure to humans may occur. Instead of a process-based approach [30] distinguished by its use of two concepts, *practices* and *interventions*, the ICRP suggested a situation-based approach to characterize exposure situations [31]. This approach replaces reference to *practices* with *planned exposure situations* and replaces *interventions* with *emergency exposure situations* and *existing exposure situations* [9]. At present, the ICRP considers three major exposure situations [31]:

1. *Planned exposure situations* are situations involving the deliberate introduction and operation of radioactive sources.
2. *Emergency exposure situations* are situations that may occur during the operation of a planned situation, or from a malicious act, or from any other unexpected situation, and require urgent action to avoid or reduce undesirable consequences.
3. *Existing exposure situations* are situations that already exist when a decision on control has to be taken, including prolonged exposure situations following emergencies.

Overall, radiation protection is based on three fundamental principles, namely, *justification*, *optimization*, and *individual dose limitation*. Within the context of remediation, the fundamental principles can be formulated as justification of remediation, optimization of remedial actions, and restriction of residual individual doses [9]. Together, they provide a basic framework for the application of remediation in long-term exposure situations.

Any remediation should be justified and remedial measures shall do *more good than harm* and shall provide optimized arrangements to maximize the net benefit to society [1]. Thus, the total benefit of reduction of exposure through remediation must be greater than any detriment that the remediation may cause. Overall, remediation shall [1]:

- (a) Reduce the doses to individuals or groups of individuals being exposed
- (b) Avert doses to individuals or groups of individuals that are likely to arise in the future
- (c) Prevent or reduce environmental impacts from the radionuclides present in the contaminated area

Justification of remediation shall go far beyond the scope of radiological protection and can include economic, political, environmental, social, and psychological facets. Justification of remediation should be based on cumulative benefits and impacts of individual remedial actions composing the remediation strategy. Some individually justified actions may be available, but may not provide a net benefit when considered as an overall strategy because, for example, collectively they bring too much social disruption for the considered exposed population as a whole, or they are too complex to manage. Conversely, a single protective action may not be justified alone, but may contribute to an overall net benefit when included as part of a protection strategy [9].

The responsibility for ensuring an overall benefit to society as well as to individuals when the population is allowed to continue living in contaminated territories is with governments or national authorities. At the same time, the current approach also emphasizes an important role to the individuals concerned. In existing exposure situations, justification should be considered for all remedial actions that may be included in a remediation strategy: those implemented centrally and locally by authorities, experts, and professionals and those directly implemented by the exposed individuals as self-help protective actions with the support of the authorities. The remediation strategy defined by the authorities should also take into account the ability of affected individuals to take self-help initiatives. However, as self-help protective actions are implemented – and thus largely decided – by the inhabitants themselves, they must be properly informed and even trained (to use the means and equipments

provided by the authorities) to take decisions concerning their own protection with a net benefit.

The process of optimization of protection is intended for application to those situations for which the implementation of protection strategies has been justified. Thus, if justified, remediation of contaminated area should be optimized. The level of protection to be achieved by the remediation should be the best under the prevailing circumstances, maximizing the margin of benefit over harm.

Owing to its judgmental nature, there is a strong need for transparency of the optimization process. The decision-making process shall provide for the involvement of a wide range of interested parties in the definition, implementation, and verification of remediation programs and for regular public information exchange on the implementation of these programs [1]. All the data, parameters, assumptions, and the values that enter into the process should be presented and defined very clearly. This transparency assumes that all relevant information is provided to the involved parties, and that the traceability of the decision-making process is documented properly, aiming for an informed decision [31].

The optimization principle should be applied in the same manner in all exposure situations. Restrictions are applied with respect to the permitted dose to a nominal individual (or *representative person*). The representative person is an individual, who will almost always be a hypothetical construct, receives a dose that is representative of the more highly exposed individuals in the population [32]. The representative person is equivalent to, and replaces, the average member of the critical group recommended previously by the ICRP [33–35].

Options which result in the *representative* person receiving a greater dose than the restrictions permit should be rejected at the planning stage. Importantly, these dose restrictions are applied prospectively, as with optimization as a whole. If, following the implementation of an optimized protection strategy, it is subsequently shown that the permitted dose for the representative person has been exceeded, the reasons for this should be investigated but the exceeding alone should not necessarily encourage remedial action.

To avoid inadequate outputs of the optimization procedure, restrictions should be placed on the

potential doses or risks to individuals who remain in contaminated territory or implement remedial actions.

The main factors to be considered for setting the reference levels for existing exposure situations are the feasibility of controlling the situation, and the past experience with the management of similar situations. In most existing exposure situations, there is a desire from the exposed individual as well as from the authorities to reduce exposures to levels that are close or similar to situations considered as “normal.” This applies particularly to situations of exposures from material resulting from human actions.

The experience gained in the aftermath of the Chernobyl accident has shown that nontechnical factors can affect remediation policies enormously [36]. The ICRP recommendations are based on objective assessments of the health risks associated with exposure levels and on radiological protection attributes of various exposure situations. However, members of the public (and sometimes their political representatives) may have personal views on radiation risks and feasibility of remedial actions intended for dose reductions. This can lead to different perceived needs for intervention and the extent of remediation, depending on the origin of the exposure. The public concerns are normally stronger when the environment is contaminated by artificial radionuclides than when these are of natural origin. An elevated exposure due to natural radiation sources are often ignored by the public, while relatively minor exposures to artificial long-lived radioactive residues are often a source of concern and sometimes lead to a request in remediation even it is not justifiable in terms of radiological protection standards.

As the long-term objective for existing exposure situations is *to reduce exposures to levels that are close or similar to situations considered as normal*, the ICRP recommends selecting the reference level for the optimization of protection of individuals living in contaminated territories in the lower part of the 1–20 mSv/year band [31].

The reference level represents the level of dose or risk above which it is judged to be inappropriate to plan to allow exposures to occur, and below which optimization of protection is implemented. Optimization is to be applied in emergency and existing exposure situations, even if the initial doses are below the defined reference levels.

Past experience has demonstrated that typical values used for constraining the optimization process in long-term postaccident situations fall in the range of the dose limit for planned exposure situations, namely, around 1 mSv/year [9]. National authority may take into account the prevailing circumstances and also usefully take account of the timing of the overall rehabilitation program to adopt intermediate reference levels to progressively improve the situation.

Although the above considerations were sometimes referred to the chronic exposures situation after the radiation accidents, the above framework is also entirely applicable for the sites of former practices and other similar exposure situations. Thus, if exposure level at the site of a former practice satisfies the reference level for all its future plausible uses, the site may be released for unrestricted use and the decommissioning phase of the practice can be terminated. If this is not feasible, the site may still be released for restricted use or released for unrestricted use after remediation. However, these restrictions and remedial actions (if any) can be considered a type of intervention since some form of institutional control will be still required [9].

Site Evaluation for Remediation Purposes

Remediation of any contaminated site starts with a site evaluation, which includes a comprehensive analysis of the factors that could affect the need of such actions and effectiveness of their application [37]. Overall, site evaluation should assist in making three major decisions. First, site evaluation should confirm that the site is contaminated to an extent requiring remediation. Second, it should be able to suggest appropriate management options if compliance with radiation safety criteria was not confirmed. And third, a site evaluation also includes a demonstration of remediation success and justification to release a site for future use [17, 28].

A site evaluation includes investigations covering [17]:

- Properties of both contaminants and the environment
- Detailed consideration of factors that could affect radiation safety at a site or of an activity that could result in a release of radioactive material and/or

could affect the dispersion of such material in the environment

- Population and access issues relevant to safety (e.g., feasibility of evacuation, location of people and resources)

To achieve these tasks, information on the site under evaluation should be made available and include data that is reliable and sufficient for making proper decisions.

If, after discovery of contamination, a radiological scan of a site shows high levels of radioactivity in some areas, then a grid survey using properly calibrated equipment is carried out. The property boundaries and pathways of the natural drainage system should be given special attention. This survey should also identify whether dose limits for members of the public or workers using a site might be exceeded under some circumstances within the boundaries of the site and to what extent [28].

If the need for application of countermeasures in the acute stage of the emergency or remediation was confirmed, a detailed site inspection needs to be undertaken. The aim of this survey is to determine the spatial extent of the problem, and to gather sufficient information to allow a full dose assessment for potential site occupiers and off-site populations. Any radiological threats to the environment from the existing situation, or from possible remediation actions, also have to be identified.

To achieve these objectives, an optimized sampling (measurement) program is required. The development of such program commences with the determination of baselines and comprises of a review of the history of all nuclear activities at the site. Alternatively, the presence of contamination has to be identified by measurements and expert judgment if no other information is available.

If a consideration of the results of this characterization leads to a decision to undertake protective actions, then detailed investigations are necessary to enable an optimized remediation strategy to be constructed. In addition, characterization of any waste stream may be necessary before the transport and disposal options for the waste are approved.

One final stage is required in the case of remediation, namely, a final survey (a compliance monitoring) to demonstrate that the objectives of remediation were successfully achieved. It also ensures that

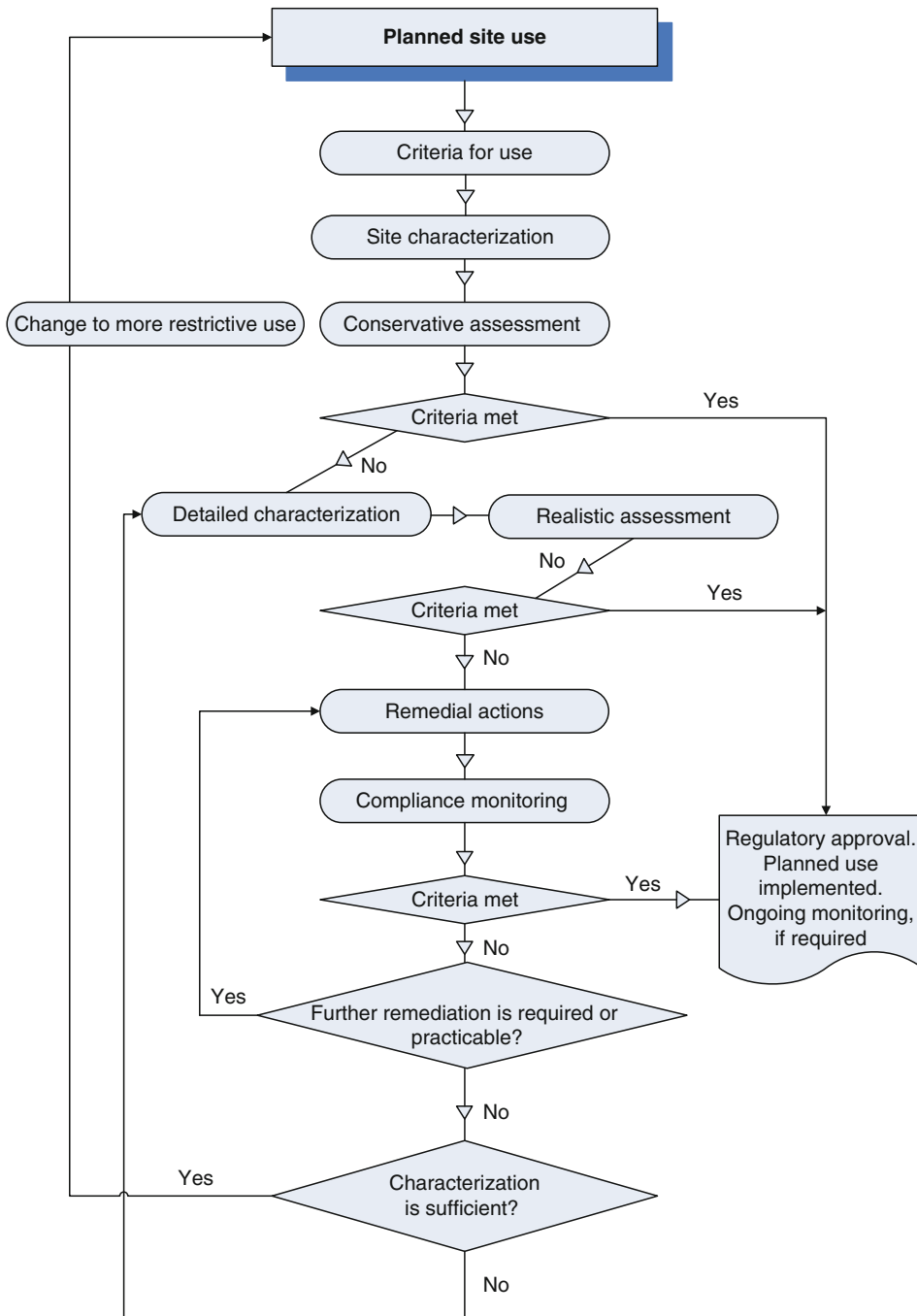
(1) radionuclide dispersion and migration will not have deleterious effects on the population and (2) the environment and the site can be reused following remediation. At that stage, compliance of residual concentrations of radionuclides in the environment with the acceptable levels must be verified and any areas with residual contamination must be identified and the nature, quantity, and distribution of the radioactivity determined. An example of the procedure, demonstrating interactions between a site remediation and evaluation within one generalized process is outlined in Fig. 1.

Management Options for the Application of Countermeasures and Remediation for Different Environments

The main objective of a management option as a countermeasure or as part of a remediation strategy is to reduce or prevent doses to humans. Although inhalation of resuspended material is a potential pathway, the doses in many situations arise largely from either the consumption of contaminated foodstuffs and water via ingestion or from exposure to contaminated surfaces in an inhabited area via external irradiation. A secondary objective is to provide reassurance to consumers and people living in contaminated areas and thereby maintain public confidence.

Countermeasures can be implemented at different phases of an emergency response extending from before the radioactive contamination is actually deposited and continuing for a varying amount of time depending on the characteristics of the accident or incident. The focus in the discussion here is on management options used after the acute phase of an accident and during remediation although countermeasures are mentioned to some extent.

The environmental behavior and mobility of radioactive contamination, and thus the potential severity of an emergency or existing exposure situation in terms of the predicted dose received by humans depends greatly on (1) the composition and physicochemical form of the radionuclides present and (2) the characteristics of the ecosystem affected by the contamination and how it is used by humans [38–43]. Thus, many different factors influence the selection of which management options are most appropriate to apply and when. When developing a remediation strategy for the long-term



Environmental Countermeasures and Restoration. Figure 1

Key stages in remediation and release of contaminated sites (Adopted from [4])

recovery of contaminated areas, some of the most important factors to consider are described below.

Factors and Processes Influencing Management Options

Management options are largely designed to reduce (1) the ingestion dose from the consumption of contaminated foodstuffs and drinking water, (2) external dose rates from deposited radionuclides, and (3) inhalation dose from resuspended material. A wide range of factors can influence the administration, impact, and consequence of the use of various management options. Some of these factors are considered briefly below and others in more detail in subsequent sections.

Temporal Factors These include the time of year that the contamination happened, the variation in activity concentrations of the deposited radionuclides with time and their rate of movement in the environment. In the short term, prompt countermeasures may be necessary such as housing dairy cattle if they are grazing outdoors at the time of deposition. However, many management options take time to organize and implement. Where management options involve significant and/or irreversible actions, then stakeholder consultation is advisable if time permits, but this may often only be possible in the recovery phase of an accident when defining a remedial strategy.

Spatial Factors The size of the area affected, extent of contamination present, and land use will influence the choice of appropriate management options can be used. In heavily contaminated areas, priority has initially to be given to the direct protection of the population rather than options for agricultural production. In other less badly affected areas, restrictions on the entry of food into the food chain may be required or management options to reduce radioactive contamination in food products. Finally, adjacently, relatively unaffected areas will also need to be monitored in case of indirect effects and to maintain public confidence. The transfer of radionuclides may vary with factors such as soil type. In remediation situations, this may mean that less heavily contaminated areas outside the major zone of concern with sensitive soil types allowing higher radionuclide uptake into plants are actually producing more highly

contaminated food products than those areas which received higher deposition [44–46].

Effectiveness of Management Options Information on effectiveness is generally presented as a percentage reduction in activity concentration in the target medium (i.e., soils, crops, building surfaces, roads) after implementation. There are some management options (such as live monitoring) which do not directly reduce doses but increase the effectiveness of other options and may also provide reassurance. Effectiveness is influenced by many factors including when it is applied relative to the time of deposition, application rates, duration of treatments, the radionuclide deposited and its form, biological half-life in animals, effective half-life in different ecosystems, timeliness of implementation, compliance, and acceptability.

Radiological Impact (Protection of Implementers)

A radiological impact can occur due to incremental doses received by people who implement management options during procedures that are not part of normal practices. The magnitude of the doses received will depend on many variables such as the type of radionuclide released and what becomes contaminated, the exposure pathways (external, inhalation, ingestion, irradiation of the skin) resulting from carrying out the management options (including disposal of wastes) and length of exposure time. Doses to implementers are controllable so the ICRP principles of protection apply, namely, justification, optimization and, dose/risk limitation.

Economic Cost Many factors influence the monetary cost of implementing management options [47]. Direct costs which are more obvious are the cost of implementation, labor, consumables, equipment and transport, and the cost of handling wastes and carrying out environmental remediation. But there are other less often considered costs such as those incurred through loss of production, loss of retail sales, closure of businesses, and the provision of temporary premises. Indirect costs are often equally important, although they are inherently more difficult to quantify. Examples are loss of market share, regional impact on tourism, which could extend to area adjacent to those which are contaminated, and the impact on society when people cannot return to normal lifestyles.

Importantly, the option not to carry out any management options can also be a valid alternative. If authorities decide that countermeasures or remediation is not required, then it is important that this decision is accompanied by a monitoring strategy aimed at reassuring the local population. This option should be considered if measurements and assessments indicate that the doses to people living in the area would be low. No judgment is made here on what would constitute a low dose. Other factors could influence the decision not to implement any remediation, such as availability of limited resources or a very large area being affected.

A large number of management options have been developed and applied, especially since the Chernobyl accident. In the 1990s, initial efforts were made in the STRATEGY project to synthesize the information on available management options for the early to late phase after an accident, and to critically evaluate whether they were generally applicable and useful to areas other than those in which they had been applied [48]. The management options can be subdivided according to the different categories discussed below; some management options in one category are also applicable to others.

The most important generally applicable management option is the imposition of food bans to prevent contaminated food products from entering the food chain. This is a highly effective, but expensive option and would be an automatically imposed countermeasure due to regulatory procedures after accidents if contamination levels were predicted to exceed the relevant action levels. Alternatively, it may be considered appropriate in certain circumstances to raise action levels. For example, this was done in Scandinavia after the Chernobyl accident to protect the Saami way of life.

Countermeasure and Remedial Actions for Agricultural Food Chains

The agricultural food chain supplies the majority of food to humans. Therefore the application of management options to this exposure pathway is a critical part of any emergency response or remediation strategy. In food production systems, management options are targeted to various media and contamination pathways including soil, crops, livestock, and other animal products. They are not only aimed at addressing health

concerns, but also a wide range of other issues such as the local economy, societal, and ethical concerns, and disposal of wastes.

In the pre-deposition phase, responses will obviously be dependent on the extent and quality of information available. Potential countermeasures include the prevention of contamination reaching food products such as switching off ventilation or irrigation systems in greenhouses, etc., or covering foodstuffs to prevent surface contamination. Livestock can also be sheltered from the contamination by being housed.

Agricultural food chain management options in the early to late phase can be divided into two types, namely, intervention along the soil-crop pathway and in animal production systems. The former focuses on reducing the contamination of crops including cereals, industrial crops, vegetables, and horticultural crops, food plants for domestic animals, and fruit. It is therefore largely aimed at reducing dose to humans. A management option which applies to both groups is the selection of alternative land uses whereby contaminated land may be used for nonfood production, such as flax, bioenergy, or biofuel crops, or for the production of leather and wool. Land may also be used for forestry, but this is unlikely to be the preferred option. The various available management options have one or more of the following aims:

- To remove contamination completely, usually through topsoil removal
- To reduce soil-to-plant transfer of radionuclides through a variety of techniques including deep, shallow, or skim and burial plowing, and soil treatment to reduce radionuclide transfer from soil to plants
- To reduce radionuclide contents at the food processing stage through chemical procedures or dilution.

If fodder plants are too highly contaminated for direct human consumption, they can still be fed to agricultural animals in the early stages of growth as long as there is sufficient time for decontamination thereafter, prior to slaughter.

Intervention in animal production systems is largely aimed at reducing the ingestion dose from meat, milk, and other dairy products, offal, eggs, and

other foodstuffs derived from agricultural animals [45, 49–52]. They generally involve:

- Decontamination of animals by feeding the animals uncontaminated feeds; the efficiency of “clean feeding” depends on the biological half-life of each radionuclide in the muscle or other products
- Reduce ingestion of contaminated feed by selective grazing regimes
- Reduce gut uptake of radionuclides by the addition of (1) specific radionuclide binders to the animal feed such as ammonium hexacyanoferrate for radiocaesium and (2) stable analogs such as Ca to suppress radiostrontium absorption
- Slaughtering of dairy livestock, which may be necessary if uncontaminated feed is not available (or other countermeasures are not possible); lactation might be suppressed in conjunction with this option
- Manipulating slaughter times to minimize the activity concentrations, by allowing physical decay of short-lived radionuclides and allowing longer finishing periods (sometimes with clean feed) to ensure that the meat is below action levels
- Reduce radionuclide contents at the food processing stage

Many of the above management options can be combined with live monitoring to determine whether activity concentration in animals are below the action limits and/or optimize application of the management options.

Countermeasures and Remedial Actions for Domestic and Free Foods from Extensive Ecosystems Including Forests

Extensive ecosystems include unimproved meadows (such as in mountainous areas or forest clearings) used for grazing livestock as well as forested areas. For grazing animals, some of the management options listed above may be applicable, although some animals in these areas are handled infrequently, so many of the most effective measures such as clean feeding are difficult to implement.

Forested areas are used by humans both for forestry, recreation, and as sources of wild foods. Dose pathways for forests may include external exposure from the

forest floor, contamination in trees, handling of forestry material, and industrial production based on contaminated wood. Internal exposure arises from consumption of forest foods such as mushrooms, nuts, and berries. Finally inhalation of radionuclides can occur following forest fires or combustion of contaminated wood [53].

The reduction of some of the doses occurring in forests soon after deposition of radionuclides requires measures aimed at the atmospheric deposition process itself and are difficult to implement, so restricting access to the contaminated forest areas may be the most important short-term countermeasure. Over longer time periods, other available management options include selective harvesting to avoid the most contaminated wild foods, preventing forest fires, providing advice on use of wood ash and reducing soil-plant uptake.

Countermeasures and Remedial Actions for Aquatic Ecosystems

Aquatic ecosystems include lakes, rivers, groundwater, and marine waters, each of which has contrasting characteristics due to hydrological, hydrochemical, or morphological factors. The need for application of management options for aquatic ecosystems is, therefore, largely dependent on site-specific parameters which often severely constrain what can be done. Furthermore, management option application in contaminated waters may be expensive, including large engineering costs so cost–benefit analysis is particularly important.

In general, population doses from aquatic pathways are often less than from terrestrial pathways depending on food habits. The main dose pathways from aquatic sources are from their use as drinking water supplies and for irrigation, and as a source of aquatic foodstuffs. If lakes used in power production, which may alter lake volume, are contaminated then the occasional exposure of contaminated bottom sediments can lead to the risk of contaminated sediments becoming airborne. Due to the self-shielding of water, external doses from recreational use of contaminated lakes and rivers can often be relatively low.

Intervention in aquatic systems involves reducing:

- Contamination reaching the water body
- Uptake by fish
- Ingestion of contaminated feed

Management options for drinking water are targeted at public and private water supplies and can be divided into two main groups:

- Provide clean water through alternative supplies and changes to the source of the supply
- Treatment of water (blending and/or purification)

Many of the above groups of management options are more effective if combined with dietary advice, the provision of monitoring kits, food labeling, and compensation schemes.

Countermeasures and Remedial Actions for Inhabited Areas

The urban ecosystem is where most people live and constitutes a wide variety of different structures and natural surfaces such as grassed areas, paved areas, exterior walls, roof paving, trees, shrubs, and interior surfaces that can all become contaminated and contribute to dose. The relative magnitude of the dose contribution from these different structures and surfaces, and thus the specific needs for management options depends strongly on the release scenario. Clean-up may result in large volumes of contaminated material requiring disposal.

For contaminated surfaces there is a need to adequately identify and quantify the problem, including the type and spatial distribution of the radioactive contamination. This is because inadequate and incomplete application of management options which do not take account of the overall spatial and vertical distribution of radioactive contamination can be much less effective than expected when correctly applied [54].

Although most remediation techniques for inhabited areas are designed to reduce external dose [55, 56], non-acute inhalation doses received from resuspended contamination can be substantial. To reduce this pathway, techniques to fix contamination to the surface to suppress dust can be used in a later recovery phase [57].

Management options for inhabited areas comprise two main groups that (1) shield people from the contamination (shielding options) and (2) remove contamination (removal options also called decontamination or clean-up options). They may be further classified according to the type of target surface

(i.e., buildings, road and paved areas, soils and grass, trees and shrubs).

Shielding options can be used to reduce both external exposure and the intake of contaminated material, and are effective in providing protection against both exposure pathways [58]. The effectiveness of a shielding option is defined as the reduction in the external dose rate from a surface (e.g., buildings, paved surfaces, grass, soil, and shrubs), generally expressed as a percentage, after the implementation of the option. If the primary aim is to reduce external exposure, shielding materials can be placed between the contamination and people (burial and covering of objects). The use of shielding materials is potentially a particularly effective option for short-lived radionuclides emitting alpha or beta radiation. Some more permanent shielding options, such as burial of contaminated material or the permanent relocation of people from a contaminated area are also effective for long-lived radionuclides and gamma-emitting radionuclides. There are three main types of shielding options:

- Burial and covering of objects including the use of clean topsoil or sand in gardens and other open areas and the turning of paving slabs, triple digging, and plowing in park areas
- Restricting people from gaining access to contaminated areas or objects, or relocating people from the area
- Fixing of contamination to the surface and restricting its mobility

Removal options involve the decontamination or clean-up of contaminated surfaces and objects. The effectiveness of a removal option is defined as the ratio of the activity initially present on a specific surface (e.g., buildings, paved surfaces, grass, soil, and shrubs) to that remaining after implementing the option. This ratio is usually called the decontamination factor (DF).

Surface treatments for buildings and infrastructure include road planning, vacuum sweeping of roads and walkways, high-pressure water hosing of walls and/or roofs, sandblasting of walls, ammonium treatment of walls, roof cleaning by cleaning device, intensive indoor surface cleaning, and firehosing of roads and walkways. One of the main disadvantages of these management options is that contaminated waste material is often produced in large quantities. Also,

decontamination techniques such as high-pressure hosing and sandblasting may damage old or poorly maintained brick or stone buildings.

Treatment of the gardens and parks can also be considered and available management options include snow removal, topsoil removal, lawn mowing and pruning, or removal of trees and shrubs.

Waste Disposal Options

Waste disposal options have been summarized most recently by Nisbet et al. [58, 59] in Recovery Handbooks and are summarized briefly here. For food production systems, considerable amounts of biodegradable waste can arise if food bans are implemented. Since the restrictions are based on statutory requirements they have to be carried out at short notice and appropriate routes of disposal need to be identified in advance of future accidents or incidents.

However, several management options also produce contaminated by-products and routes for their disposal must be considered. Contaminated produce that might require disposal include crops and by-products from processing of grass products (fresh grass, silage, hay), milk and by-products from processing, whole animal carcasses and meat, and soil.

Waste treatment options that can be carried out on-site include composting, land spreading of milk and/or slurry, and plowing in of a standing crop. Off-site options include biological treatment (digestion) of milk, burial of carcasses, disposal of contaminated milk to sea, incineration of crops, landfill, processing and storage of milk products for disposal, and rendering of animal carcasses (which converts waste animal tissue into stable, value-added materials).

The contamination of an inhabited area generates both solid and liquid waste regardless of whether any countermeasures are applied. In inhabited areas, contaminated items may need to be removed and facilities such as those for water treatment may become contaminated by waste water from management options such as firehosing buildings and by contaminated precipitation and runoff. The categories of radioactive waste are (1) contaminated waste (refuse) and goods, (2) solid and liquid waste from clean-up of the contaminated area and (3) waste water from precipitation and natural run-off. It is therefore important to

consider the impact of the contaminated waste on the public, workers handling the waste, the environment, and normal waste disposal practices.

Waste disposal schemes for solid contaminated waste should preferably be planned for in advance, with potential disposal sites and resources identified if possible [54]. The selection of suitable disposal sites needs to take account of the proximity, cost, hydrogeological characteristics of the site, geological stability, and future land use.

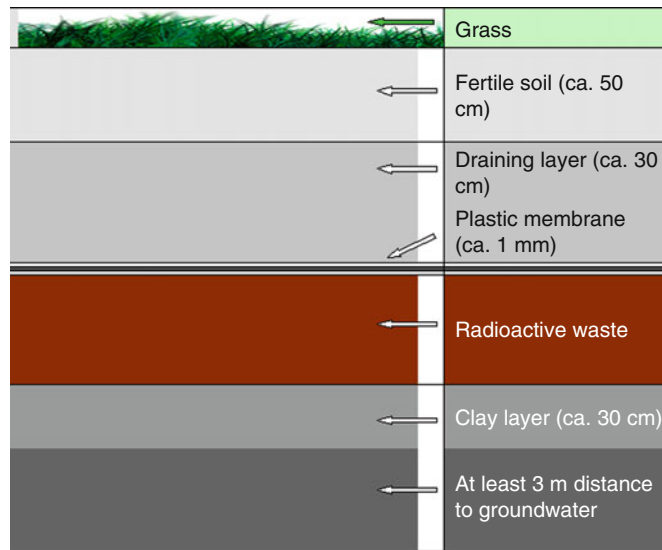
Construction of a repository (Fig. 2) should consider a number of other factors:

- There should be adequate shielding so that the contamination in the repository does not contribute significantly to the external dose rate in the area.
- The ecological characteristics of the area should be assessed and intrusion by humans, animals, and plant roots needs to be prevented, usually by using suitable protective layers.
- Migration of radionuclides from the waste with rainwater must be prevented, usually by draining the rainwater off in a gravel layer on top of the waste layer.
- The repository should be protected against erosion.
- The acceptability of constructing repositories needs to be assessed.
- The repository may be equipped with a drainpipe at the bottom for inspection of the radionuclide contents in any water passing through the construction.
- The design should be simple, inexpensive, and possible to construct quickly.

Secondary (or Side) Effects of the Application of Management Options

Environmental impacts occurring due to the application of management options can be positive or negative, direct or indirect, and can lead to social disruption or damage to the environment [60]. Direct environmental impacts can include changes in biodiversity, soil fertility and structure, and enhanced soil erosion as well as pollution of air and aquatic ecosystems. Indirect effects may be less immediately obvious, but could include restrictions to an individual ability to follow their selected lifestyle.

The Chernobyl Forum, based on the related studies in areas affected by the Chernobyl accident, reported



Environmental Countermeasures and Restoration. Figure 2

The principles of a simple ground repository for radioactively contaminated remediation waste (e.g., soil), as recommended by Junker et al. [53]

that some management options led to detrimental social and economic effects, particularly a breakdown in the social infrastructure [61] which could also be considered amongst the most serious consequences of the accident. Nevertheless, remediation can also include options to restore ecosystems, and secure livelihoods and the social structure of affected populations, or stabilize the economic situation [62].

Side or secondary effects of management options can be very diverse. If all access by the public is restricted and access for forestry is limited, major losses to the population would include a lack of locally collectible: wood for industrial purposes and for heating; berries and mushrooms; forest fodder for dairy cows and firewood for rural populations. The cost of forestry is also increased due to limitations on the time which forest workers can spend in contaminated forests which, in turn, can lead to deterioration in the maintenance of forests with negative ecological consequences. Application of restrictive management options such as limited access to forest in the areas affected by the Chernobyl accident had negative psychological and sociological consequences. The economic losses were also large [60, 62].

Implementation of soil-based agricultural actions can change soil properties or decrease air and water

quality, while land use-based remedial options can change the landscape and its use, impacting the economic or recreation leisure value of an area [12, 48, 60].

Societal and Ethical Aspects

Overall, the primary factors and effects described above led to many ethical issues including assessment of the consequences of remedial actions implementation for future generations, sustainability, and comparing harm to the environment to benefits to humans. When both radiological factors and social aspects are taken into account, better acceptability of countermeasures by the public can be achieved [12, 61]. Thus, the impact on individual and communities needs to be considered. Local stakeholders need to be consulted in the identification of problems and their solution.

Some management options may have a negative impact on society by causing disruption (e.g., through restricting access or activities), anxiety, and stress (e.g., by causing panic, upheaval), and stigma (e.g., by affecting businesses or tourism). These need to be balanced against positive impacts through provision of reassurance and improvements to living conditions. There are also ethical issues to be addressed when selecting management

options, including: provision of self-help options that reinforce liberty and dignity, the distribution of doses over space and time and between different members of the community, animal welfare, environmental risk, and consequences for future generations [62].

The ethical acceptability of remediation is highly dependent on the ecological status of the area and the degree to which the actions diverge from usual practice. In most cases, environmental legislation must also be considered [63].

Studies on effectiveness of remediation implementation in many areas affected by the Chernobyl accident have identified a variety of factors that had a highly detrimental impact on the economic and social activities on contaminated areas. Practically all remedial options were too expensive to be competitive at the market, and required large subsidies from the state to support economic activity in the affected regions. In most cases, remedial actions imposed a stigma on the areas with contaminated environments [62]. Similar psychological effects have also been observed in some other European countries [64]. This constrained or prevented export of even high-quality products to nonaffected areas, constrained economic and social development of those regions, and promoted migration of some of the working-age population to areas with lower contamination levels. As a consequence, economies and social structures in affected communities deteriorated, alongside an apparent increase in poverty [62].

The provision of information, and how that information is communicated, will have a significant influence on how the authorities manage the situation, on society's response to the problem and on the overall success of the recovery strategy. Maintaining public confidence is paramount. Trust is easily lost and difficult to develop and/or regain so it is important to create a framework for information and communication strategies under noncrisis conditions with the involvement of stakeholders. The type of information disseminated should be targeted to meet a variety of needs. The form of communication should be adapted to different levels of understanding, circumstances and to address the relevant issues. This should be implemented at the same time as the development of restoration strategies.

Management option implementation is generally the responsibility of the authorities and carried out

by designated personnel. Nevertheless, self-help implemented by the affected population can also be considered. Some of the simple management options that do not require any specific skills or experience to implement (such as lawn mowing, digging) are potentially suitable for implementation by affected inhabitants themselves [65]. This could, for instance, be beneficial in providing an extra labor force in situations where large areas need to be treated quickly. The involvement of affected persons in actions to improve their own situation can be psychologically beneficial and give a better feeling of control of the situation, which also prevents undue anxiety. Health and safety issues need to be carefully considered for people carrying out such "self-help" options, need careful communication and to be conducted with adequate supervision of individuals to ensure they are implemented correctly.

The Compendium of Management Options Datasheets

The concept of systematically describing all the key factors which need to be considered in the application of management options in a datasheet format was developed in the mid-1990s within the STRATEGY project [41, 48, 55, 66]. Standardized datasheets, mainly based on the Chernobyl experience, capturing key information required to carry out the management options and evaluate their feasibility, applicability, and practicality were developed to aid the decision-making process and to ensure that potentially important issues are not overlooked. They also described the resources needed. The information in the datasheets was intended to be generally applicable for use in different European countries. The rural and waste templates are being used by the joint FAO/IAEA division to be adapted to other climate types and to update the IAEA Handbook of parameter values for radionuclide transfer in the temperate environments 364 [67].

Stakeholder participation is an important mechanism to explore these additional benefits or disadvantages to the use of management options, and is an essential step in developing a decision framework which avoids problems previously experienced in emergency management [68]. Close liaison with the FARMING stakeholder network facilitated evaluation

of management options datasheets for rural ecosystems; while for other datasheets, stakeholders were consulted in small groups or individually. Stakeholder opinion suggested that some management options were as likely to be rejected on socio-ethical grounds as on technical and economic grounds. Rejection of specific management options can be expected to show site, context, and national differences.

The datasheets developed in STRATEGY have been gradually improved during the EURANOS project [69] with more detail provided and the scope widened to provide information representing more diverse conditions and for wider range of radiological hazards. Sources of contamination considered in the handbooks include nuclear accidents and radiological dispersion devices and the most relevant radionuclides are included. In addition, non-radiological considerations have been incorporated in the later generations of datasheets [48, 57], introducing viewpoints from experts on social, ethical, and economical sciences [68], and also considering documented reactions of the public and other participants during management option trials in the Former Soviet Union.

The most recent compendium of datasheets is integrated into decision-aiding handbooks for contaminated food production areas [58, 59] and inhabited areas [57]. These handbooks contain advice on the development of optimized strategies for remediation under specific conditions using a series of diagrams, tables, and decision-charts, which guide the user through a series of eliminations of unsuitable options. Site-specific factors need to be taken into account in this process, and it is recommended in the handbooks that they be customized and tailored for different users. The importance of adequately involving people in decisions that affect them is stressed.

The most recent revision of the datasheet template for food [58] and inhabited areas [55, 57] includes a wide range of criteria describing different aspects of each countermeasure including the objective and benefits, the target of the measure and the radionuclides targeted, constraints (e.g., legal and social), effectiveness, feasibility (equipment, skills, infrastructure needed), waste produced, doses to implements, costs, side effects of implementation, and stakeholder opinion of the measure.

The datasheets provide a comprehensive overview of management options [69] which are intended to be useful to “middle-level” decision makers who have to assess possible countermeasure options. They are not a comprehensive set of instructions, but provide a template to adapt to local conditions/languages. Their applicability will reduce across the world as climate, culture, production systems, and drinking water supply deviate from those currently encountered in Europe.

Countermeasures and Remediation in Areas Affected by Radiation Accidents

Various radiation accidents requiring an organized emergency response have resulted from a variety of different factors, including human error (Chernobyl accident), violations in procedures at a radioactive waste tank (Kyshtym accident), the crash of aircrafts and subsequent loss of nuclear weapons (e.g., Palamares, Spain and Thule, Greenland) and the loss or theft of sealed radioactive sources used for medical purposes (Goiânia, Brazil). Contamination scenarios and the radionuclides released differed in each case, thus requiring a variety of different countermeasures and remediation strategies.

Kyshtym Accident

An accident at a nuclear facility producing weapon grade plutonium in Chelyabinsk-40 (Soviet Union) led to one of the largest radioactive releases into the environment. Sr-90 was the most important dose forming radionuclide in the release and required extensive countermeasures and remediation. The resulting East Urals Radioactive Trace (EURT) was defined as the area where the ^{90}Sr contamination density at the time of the accident was above 3.7 kBq m^{-2} , a value which was twice that of global fallout of ^{90}Sr from nuclear tests. In the EURT area ($23,000 \text{ km}^2$) there were 217 settlements with 270,000 inhabitants [70–74]. Within the EURT, a contamination density of 74 kBq m^{-2} for ^{90}Sr was used as the criterion to identify areas requiring management options [74]. Heavily contaminated areas were in a narrow band, 4.5–6 km wide and 105 km long covering an area of $1,000 \text{ km}^2$ (c. 5% of the EURT area). All economic activity within the contamination isolines of 74–150 kBq m^{-2} of ^{90}Sr was suspended

(resulting in a restriction zone of 700 km²). In 1959, this area included 106,000 ha of which agricultural lands, forests, and lakes accounted for 54%, 36%, and 9%, respectively [70].

Agriculture The Kyshtym accident was the first case where vast areas received countermeasures soon after the accident and remediation activities in the long term. In the first phase, the major contributor to the dose was the internal exposure from ¹⁴⁴Ce uptake by humans with foodstuffs [73]. The maximum concentration of radionuclides in agricultural products on land closest to source areas (up to 20 km) reached 10–10,000 kBqkg⁻¹, ¹⁴⁴Ce and ⁹⁵Zr being the main contributors to these values except for milk where ⁹⁰Sr comprised a mean of 70% of the activity concentration present [25]. The radionuclide concentration in all environmental media, including farm products, declined with time reflecting the ecological half-life of the contaminated territory. After 5–8 years (after the ⁹⁵Zr, ¹⁰⁶Ru, and ¹⁴⁴Ce had largely decayed away), only ⁹⁰Sr, and to a very small extent ¹³⁷Cs, were relevant for determining radiation dose to humans [26]. In response, soil-based management options were a focus of the remediation strategy adopted (Table 1).

For some small areas with very high contamination levels, the decontamination of land by removal of the upper soil layer (5–10 cm depth) using bulldozers, graders, or scrapers with subsequent burial in special facilities was very effective (Table 2). This action was widely employed despite the high operational cost. Plowing rapidly after deposition greatly reduced external dose rate. Plowing of soil with the burial of the most contaminated upper layer (deep plowing) soon after the accident was one of the most effective methods that brought about a threefold reduction in external doses [74]. Plowing and burial were especially effective on chernozem soils with a thick humus horizon (up to 0.5 m); on low-fertility soils it was less effective. The decontamination effect from deep plowing was enhanced by adding mineral fertilizers at balanced rates (according to plant demand for mineral nutrients) and liming where appropriate. These widely used techniques provided up to a tenfold reduction in ⁹⁰Sr concentrations in plants compared with conventional plowing [26, 74]. Normal plowing (20–25 cm of soil)

Environmental Countermeasures and Restoration.

Table 1 Effectiveness of soil-based options for different crop types [24–26]

Treatment	Soil type	Reduction factor of ⁹⁰ Sr in root uptake
Liming:		
0.5 H _g ^a	Mineral	1.1–2.1
1 H _g		1.4–2.7
2 H _g		1.7–3.1
1 H _g	Organic	1.0–1.1
Mineral fertilizers:		
N ₉₀ P ₁₈₀ K ₉₀	Mineral	1.1–1.4
N ₆₀ P ₉₀ K ₁₂₀		1.3–1.9
Plowing:		
Plowing, depth of cultivation 20–25 cm	Mineral	1.0
Moldboard plowing, depth of cultivation 50 cm	Mineral	1.3–2.3
Plowing with turnover of upper layer	Mineral	1.6
Soil removal		5–15

^aH_g Hydrolytic acidity

Environmental Countermeasures and Restoration.

Table 2 Summary of reduction factors of soil based options used in the former Soviet Union countries after the Chernobyl accident [16, 49, 52]

Countermeasure	Reduction factor	
	¹³⁷ Cs	⁹⁰ Sr
Normal plowing (first year)	2.5–4.0	
Skim and burial plowing	8–16	
Liming	1.5–3.0	1.5–2.6
Mineral fertilizers	1.5–3.0	0.8–2.0
Organic fertilizers	1.5–2.0	1.2–1.5
Radical improvement (first application)	2.0–9.0 (2.0–3.0) ^a	1.5–3.5 (1.5–2.0)
Removal of soil (external dose)	1.5	Not relevant

^aMultiple applications

for the first time after deposition achieved a 1.1–2.4-fold decrease in dose rate of γ -radiation. A further reduction in ^{90}Sr uptake in plants (including fodder for animals) was achieved through cultivation of plants with low ^{90}Sr accumulation giving a roughly tenfold reduction in the main farm crops although the reduction for some species was up to 60-fold [26, 74].

Management options to reduce ^{90}Sr accumulation in animal products mainly consisted of the provision of fodder with low amounts of ^{90}Sr to animals (clean feeding). The most effective feeds were based on potatoes, root vegetables, and grain which had much lower ^{90}Sr concentrations compared with feeds from unimproved lands typical of preaccident farming practice. In specialized agricultural farms this clean feeding resulted in a decrease of ^{90}Sr content in meat and milk by two- to sevenfold and three- to fourfold, respectively compared with normal farms [74, 75].

A reduction in ^{90}Sr transfer to animal products was also achieved through the administration of Ca additives or fodder rich in Ca (e.g., legume crops) resulting in a tenfold decline in ^{90}Sr concentrations in milk. Sr-90 reduction in meat was also achieved by preslaughter feeding with less contaminated fodder [26, 74].

Further reductions in the ^{90}Sr concentrations in food products and, hence, the internal doses to humans, was achieved by processing raw products. The processing of milk to fermented milk products and butter, which reduces ^{90}Sr concentration in butter 20-fold compared with raw milk was widely used. Conventional processing of grain reduces ^{90}Sr concentration in the final products by two to three times. Starch, alcohol, and vegetable oil were produced “clean” from the original contaminated agricultural products after the normally used processing [26, 74, 75].

Forest and Aquatic Ecosystems The key feature in the application of environmental management options in natural ecosystems affected by the accident was the identification and implementation of action limits for forest products including berries, mushrooms, game, and wood. The maximum permissible ^{90}Sr contamination density for forested areas used for pasture and haymaking was set at 92 kBq m^{-2} and for the production of industrial wood at 3.7 MBq m^{-2} [26, 74]. Successful implementation of these options was achieved by transferring the priority right to use these lands to

specialized farms, and controlling public access to those areas.

Initial restrictions on the economic use of water bodies (fishing, use of aquatic vegetation as fodder for farm animals) were imposed on lakes in areas with ^{90}Sr density above 74 kBq m^{-2} [71]. By 1970, reutilization of lakes located in the affected area was possible for fish farming due to natural decontamination in lakes (the ^{90}Sr ecological half-life in water was 5–6 years) (Alexakhin et al. 2004 [25]). The remediation strategy was based on the presumption that lands should not be left vacant (with the caveat that occupational standards and dose limits to individuals must be observed) and their exploitation must partially compensate for damage from radioactive contamination. It was based on two main options: land use change and development of new economic activities, dependent on safe use of the environment. Land use differentiation into agriculture, forestry, and other branches of the rural economy using soil, plant, and water resources was based on the ability to produce products with permissible levels of ^{90}Sr contamination. Since farming was seriously restricted in the affected region (due to exceeding action levels), economic activities were focused on the production of industrial (nonfood) goods, development of forestry and local industry and the exploitation of peat, sand, gravel, and other mineral resources [26].

Effect of Remediation As a result of the remediation, a substantial decrease in the effective annual doses to the local population was achieved which ensured that these doses fell below those set out in the radiation safety standards present in the USSR at that time. The complex mixture of protective actions which was initially developed for areas affected by the Kyshtym accident was subsequently adopted for areas affected by the Chernobyl accident, with appropriate modifications to take account of the different circumstances [25, 70, 71].

Nuclear Weapon Accidents: Palomares Accident

Although a number of accidents involving nuclear weapons are currently known to have occurred (with different extent of reliability) (see <http://www.atomicarchive.com/Almanac/Brokenarrows.shtml>), a limited number, in particular accidents with losses of

nuclear weapons from aircraft, required remediation efforts.

In January 1966, near the south-eastern coast of Spain, two American air force aircrafts collided during air refueling. Of the four nuclear bombs that were onboard the aircraft, one was found in the Mediterranean Sea. The second bomb fell by parachute onto agricultural fields. The remaining two bombs were physically destroyed due to the explosion of their chemical detonators. Pyrophoric metallic plutonium caught fire and the resulting radioactive cloud, containing plutonium oxide, moved westward contaminating an area of 262 ha within a region of intensive farming, occupied by rural dwellings, agricultural fields, and forested areas [19].

The main remedial management options were applied to the soil. The soil was removed, placed into containers, and shipped to the US for disposal at the Savannah River plant at a Pu deposition density above 1.2 MBq m^{-2} . The total area of decontaminated soil was 2.2 ha, removed in 6,000 containers with a volume of 250 L each. Arable soil with Pu concentrations under 1.2 MBq m^{-2} (17 ha) was re-plowed to a depth of 30 cm. In rocky areas where plowing was impossible and Pu levels exceeded 120 kBq m^2 , soil was removed by hand and shipped to the USA [19].

Upon completion of mechanical decontamination, radiation monitoring of the air, soil, plants, farm animals, and the population was carried out. The major farm crops produced in the affected area were tomatoes, barley, alfalfa, maize, pepper, and fruit such as melon. Contamination levels in agricultural products were not sufficient to cause serious exposures to the public. Twenty-three years after the accident approximately 99% of the Pu still remains in the 0–5 cm surface soil layer [26].

The accident produced no serious health impacts, since doses were very low. The committed effective doses calculated for total Pu intake over 30 years amounted to 37 and $210 \text{ } \mu\text{Sv}$ for urban and agricultural workers (the most exposed population group), respectively [76].

Chernobyl Accident

The accident at the Chernobyl NPP was the most serious radiation accident in the history of nuclear energy generation. The accidental release comprised some highly

environmentally mobile radionuclides including isotopes of radioiodine, radiostrontium, and radiocaesium, which had a serious impact on agriculture [12].

More than 15,000 settlements with a population of around six million people were living in areas with a ^{137}Cs deposition density above 37 kBq m^{-2} , and were therefore officially designated as contaminated. In the areas most severely affected, with ^{137}Cs contamination densities exceeding 555 kBq m^{-2} , there were 640 settlements, with 270,000 inhabitants [61, 77, 78].

Soon after the accident, a 30-km radius exclusion zone (the CEZ) was established around the reactor. Further population relocations took place in subsequent months and years in Belarus, the Russian Federation and Ukraine. Eventually, 116,000 persons were evacuated or relocated [61] with about 50,000 cattle, 13,000 pigs, 3,300 sheep, and 700 horses [12]. In the 30-km zone, more than 20,000 agricultural and domestic animals, including cats and dogs, were slaughtered and buried. In the first few years after the accident comprehensive surveys of contaminated lands and on-site inspection of foodstuffs were carried out. In 1989–1991, people from 470 settlements outside of the 30-km zone of the Chernobyl NPP were moved together with their agricultural animals to new locations [61].

Although decontamination of some settlements was carried out within the first few years after the accident, the main challenge for remediation after the accident was to retain the capability for agricultural production in contaminated areas. This task was of paramount importance to preserve the economy and the social and psychological stability of the population [12].

Agriculture The maximum radionuclide concentrations in agricultural products occurred in the farms surrounding the Chernobyl NPP in the first few days and weeks after the accident and considerably exceeded action levels [12]. During the first few weeks after the accident ^{131}I was the main contributor to internal dose leading to significant radiation doses through milk consumption [61]. In addition, plants and animals were also contaminated with a wide range of other radionuclides, in particular, radiocaesium isotopes which were the dominant contaminant in most environmental samples and food products from June 1986.

In the first weeks after the accident, the main goal of countermeasures was to prevent the consumption of contaminated milk and green vegetables by the population, and to decrease ^{131}I concentrations in milk [16]. This was largely achieved by (1) radiation monitoring of agricultural products and the rejection of foodstuffs in which ^{131}I activity concentrations were above action levels ($3,700 \text{ Bq l}^{-1}$ at that time) and (2), processing rejected milk (mainly converting milk to storable products such as condensed or dried milk, cheese, or butter). Information on countermeasures was confined to managers and local authorities, and was not distributed in time to private farmers. As a result, application of the countermeasures was often delayed in rural settlements for privately produced milk, resulting in low countermeasure effectiveness in some areas.

The rural population around the ChNPP was the most heavily affected by both external and internal exposure pathways. The production of foodstuffs with radionuclide activity concentrations above action levels was considered by the population as evidence that the land and their lifestyle had been severely blighted. Therefore, the rapid deployment of optimized countermeasures was important not only to reduce the radiation doses but also for the social and psychological well being of the population of affected regions and helping to decrease their stress levels.

In the least contaminated zone ($37\text{--}185 \text{ kBq m}^{-2}$) farming continued based on agricultural practices suitable for the soil and climate, except for areas with peaty soils, with high radiocaesium uptake by plants, where radical improvement was implemented. In the $185\text{--}555 \text{ kBq m}^{-2}$ zone, plant production on arable soils continued without restrictions. Soils received increased doses of P and K or organic fertilizers. Natural meadows on sandy, sandy-loam, and peaty soils were subjected to radical improvement (which includes removing vegetation, plowing, liming, fertilization, and plants reseeded) with a 1.5-fold increase in the application rate of potassium and phosphorous fertilizers [12, 79] (Fig. 3). In the $555\text{--}1,480 \text{ kBq m}^{-2}$ zone, there was a full-scale remediation. The areas receiving soil based management options in the 20 years following the Chernobyl accident are shown in Fig. 4 and the effectiveness of these actions is given in Table 2.

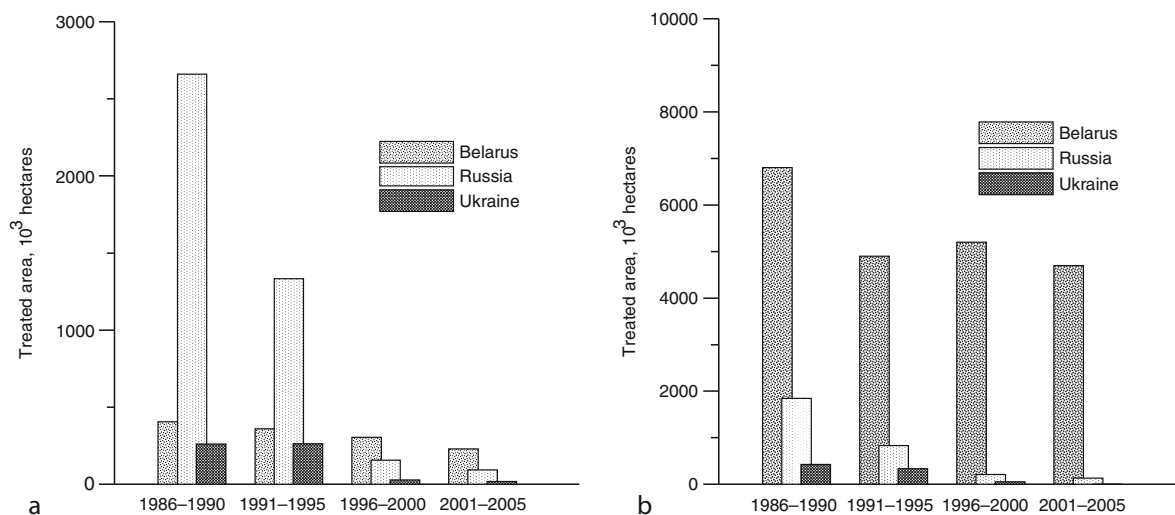
The provision of feed with low contamination levels to previously contaminated animals for an appropriate



Environmental Countermeasures and Restoration.

Figure 3

Implementation of soil based remedial actions (radical improvement) in Rivno region of Ukraine (*upper photo*) and Tula region of Russian Federation (*photo in the mid*). Lower picture shows a meadow after radical improvement (Bryansk region, Russia). (Courtesy of Russian Institute of Agricultural Radiology and Agroecology Obninsk, Russia and Ukrainian Institute of Agricultural Radiology Kiev, Ukraine)



Environmental Countermeasures and Restoration. Figure 4

Changes with time in the extent of agricultural areas treated with liming (a) and mineral fertilizers (b) as a countermeasure in the USSR/FSU countries contaminated by the Chernobyl accident [12]

period before slaughter effectively reduced contamination in meat. Such “clean feeding” was normally combined with live monitoring of animals so that if radiocaesium transferred to meat was above action levels, the animals were returned to the farm for further clean feeding (Fig. 5).

Another important countermeasure used since the beginning of the 1990 s to reduce radiocaesium contamination in animal products was administration of cesium binders (Fig. 6). The Cs binder used was a hexacyanoferrate, widely referred to as Prussian blue [80] and was administered as a powder and as rumen dwelling boli.

This application has been especially useful and effective in settlements where the grazed meadows were not suitable for radical improvement. Various methods for milk processing to butter and sour-milk products were also developed and successfully implemented during food processing.

Forest and Aquatic Ecosystems More than 40,000 km² of forested areas had ¹³⁷Cs concentrations above 37 kBq m⁻² and in Russia alone about 50,000 forestry workers and members of their families lived in these affected areas [25].

In many of the most affected areas, the role of the forest as a source of food is essential to local

communities [81]. Natural reductions in the ¹³⁷Cs concentrations in “forest products” was much slower compared with agricultural products; the effective half-lives for ¹³⁷Cs in mushrooms and berries were roughly 20 and 10 years, respectively [82, 83]. As a consequence, in the long term after the accident “forest products” (especially mushrooms) sometimes contributed more to the exposure of a population than did critical food-stuffs such as milk. This was particularly prevalent in settlements near forests [14, 39, 81].

Remedial actions in forests were mainly of a restrictive nature and the dose reduction achieved by various management options was small [84, 85]. Thus, in the affected forested areas, bans were imposed on the collection of mushrooms, berries, hunting game, and private cattle grazing on forest pastures and clearings; fire-control actions were also implemented. Management options in contaminated woods (such as the collection and removal of forest litter and leaves or reforestation) were either technologically ineffective, or economically expensive and ecologically unjustified.

The Chernobyl accident also affected many water bodies. Some 8 million people were exposed to additional radiation due to the consumption of drinking water and 32 million due to the consumption of fish and the use of water for irrigation purposes [61, 86].



Environmental Countermeasures and Restoration.
Figure 5

Live monitoring measurements of animals in the settlements of the Bryansk region Russia affected by the ChNPP accident (Courtesy of Russian Institute of Agricultural Radiology and Agroecology, Obninsk, Russia)

Two major remediation strategies were implemented in water bodies affected by the Chernobyl accident: the implementation of remedial options to restrict radionuclide distribution with water movements; and restrictions on the consumption of aquatic foodstuffs and drinking water from the most contaminated water bodies. In the contaminated catchment area different hydro-technical constructions were implemented (dikes, dams, etc., as well as special channels). Also, to limit radionuclide migration and accumulation in fish, various sorbents were added to induce radionuclide sedimentation from water masses, and



Environmental Countermeasures and Restoration.
Figure 6

Administration of the hexacyanoferrate boli to a sheep (Courtesy of Centre for Ecology & Hydrology, Lancaster Environment Centre, Lancaster, UK)

potassium fertilizers were added to reduce radionuclide accumulation in fish [87, 88] (see example in Fig. 7). These measures were rather costly and their effectiveness was generally low, especially compared with agricultural remedial options [26, 61].

Urban Ecosystems The decontamination of settlements following the Chernobyl accident was applied at a limited scale. For example, the Soviet army in 1989 decontaminated 93 rural settlements affected by the Chernobyl accident in the Bryansk region. The army units removed a layer of topsoil from areas around houses, from children gardens, and other public sites [54]. However, the dose reducing effect was only 1.1–1.5; which was much lower than expected [14]. Therefore, decontamination of inhabited areas ceased and relocation of the residents was continued in some cases. In later studies, carried out in 1997 it was shown that if soil removal is optimized in relation to the distribution of the contaminants in soil profile (e.g., using a mini-bulldozer, see Fig. 8), the external dose rate can be reduced by at least a factor of 6 [54, 89].

Effect of Remediation Countermeasures applied in the early phase of the Chernobyl accident had a considerable effect in averting doses to the affected population. However, they were only partially effective in reducing radioiodine intake via milk, because of the lack of timely information about the accident and



Environmental Countermeasures and Restoration. Figure 7

Application of potassium chloride fertilizer for reduction of ^{137}Cs concentration in the fish of the Svaytoe Lake (Belarus) (left photo) and the lake after the treatment (right photo). The fertilizers were colored red with a special neutral additive to track a potassium chloride application



Environmental Countermeasures and Restoration. Figure 8

Removal of contaminated topsoil with a "Bobcat" mini-bulldozer in the Bryansk region in 1997 [54]

advice on appropriate actions, particularly for private rural farmers [61].

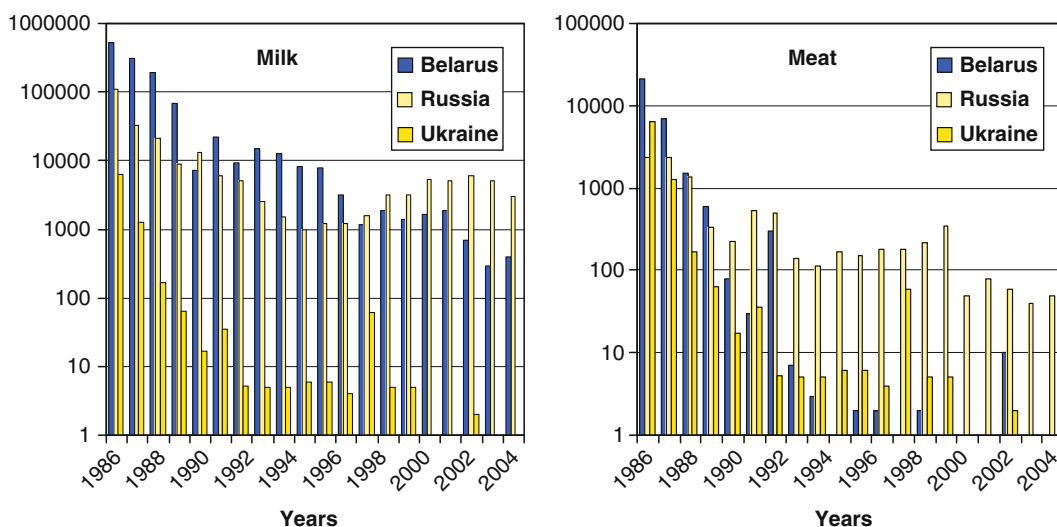
Large-scale application of a range of management options achieved a sharp decrease in the amount of animal produce with radiocaesium activity concentrations above action levels in all three countries (Fig. 9).

The contribution of internal exposure to radiation doses of the general public caused by the Chernobyl radioactive release has been significant both in the short and long term. Agricultural management options were the most efficient in reducing dose, averting 30–40% of the collective internal dose to the affected population, excluding thyroid dose [12].

The current slow, natural decrease of ^{137}Cs availability in soil means that remediation actions after the Chernobyl accident will be important for around 70 years after the accident except for those areas (such as the 30-km zone around Chernobyl NPP) which have been withdrawn from economic use and for which this time period is longer.

Goiânia Accident

A serious radiation accident occurred in September 1987 in Goiânia in Brazil. One of the hospitals in Goiânia possessed a 50.9 TBq (1,375 Ci) ^{137}Cs ($^{137}\text{CsCl}$) sealed source for therapeutic applications [2, 3, 90]. Several unemployed young people penetrated the unguarded building and stole the device. They thought that the 50-kg metallic object was made from a valuable material so they separated the source from its biological shield and opened the source container. The source was sold at a junk yard; the junkyard



Environmental Countermeasures and Restoration. Figure 9

Variation with time in amounts of milk and meat exceeding action levels in Russia (all milk and meat – collective and private), Ukraine and Belarus (only milk and meat entering processing plants) after the Chernobyl accident, tones [12]

owner considered the material to be precious because of its blue luminescence. Dispersion of the radioactive substance then occurred as people came to look at the material, some of whom then took away pieces of it. Some people took parts of the source as souvenirs; others painted their faces with the blue powder for carnival.

As a result, the source was broken into small fragments and widely distributed. Radioactive substances were dispersed by high winds over a variety of vegetable gardens, houses, and roofs of buildings. Radioactive material even reached the main river in the region, the Meia-Ponte River with ^{137}Cs found in bottom sediments 12 km from the source site. In this incident some 250 persons were exposed to external and internal irradiation; 50 of which developed symptoms of whole-body and acute local irradiation; 14 of these had marrow failure. Doses of internal exposure in affected people ranged from 0.046 to 0.97 Gy [3].

Remediation actions for the Goiânia accident were carried out with international assistance. Some 755 professionals took part in remediation of the contaminated sites and were exposed to differing extents. Doses to around 70% of the participants were below 1 mSv and the maximum dose value reached was 16 mSv [91]. Very high contamination was found in 85 houses and residents from 41 houses were

evacuated. The total area subject to radiation monitoring within the urban territory covered 67 km². In the city a γ -survey was performed by helicopter, followed by another γ -survey performed by vehicle.

The limits of contamination density chosen for the remediation actions in Goiânia (3.7 Bqcm⁻² or 37 kBq m⁻²) were the same as for the classification of the Chernobyl (1986)-affected area (areas with ^{137}Cs levels 37 kBqm⁻² or 1 Cikm⁻²). Once contaminated sites had been identified, more labor-intensive options were implemented. Various chemical solvents and abrasives were used for mechanical decontamination and radionuclide-containing materials were physically removed [92].

The identification of potential locations to store the collected radioactive materials was met with a highly negative public response. Ultimately, a site for temporary storage was chosen in the thinly populated town of Abadia de Goids (20 km outside the city of Goiânia). In the most affected areas houses were destroyed, soil was removed, and the decontaminated site was covered with concrete. In less-affected buildings, walls, floors, and roofs were decontaminated. Personal items that contained radioactive substances were designated as waste, since people refused to use them even after decontamination. In addition to houses, 45 public places including parts of streets, squares, parks, and

shops were decontaminated. By the end of December 1987, the main affected sites were decontaminated. Remediation after the Goiânia incident resulted in the bulk of the ^{137}Cs source (50.9 TBq) being localized in the 1.6-km² waste disposal area. The remediation in Goiânia thus largely comprised of the mechanical cleanup of contaminated territories, collection of various contaminated materials which were later assigned to radioactive waste, and their disposal at specially allocated storage sites [93].

Remediation of Sites Affected by Nuclear Testing

Radioactive contamination is often the result of human actions, the consequences of which were not fully understood at the time, as is the case for nuclear testing sites. Nuclear tests were conducted at a total of 16 sites, the most well known of which are the Nevada Test Site (the USA), Bikini and Enewetak (Marshall Islands), Emu, Maralinga, and Monte Bello Islands (Australia), Mururoa and Fangataufa (French Polynesia), Semipalatinsk (Kazakhstan), Novaya Zemlya (Russia), and Lop Nor (China) [6, 78, 94–97]. In total, about 29 mountains of fission yield were associated with radionuclides locally deposited at the test sites [98]. Another source of local contamination at the test sites were safety tests, in which nuclear weapons were destroyed using conventional explosives to simulate a possible accident.

Only two of these sites (Enewetak and Maralinga) have been remediated. Remediation at the other test sites was either not carried out or only partly performed, either because of low residual contamination (fractions of long-lived radionuclides in many types of nuclear explosions are very low), high financial costs, the remoteness of the site from populated areas or other political reasons. Even when a full remediation strategy was not applied at these sites, some remedial actions were usually conducted, especially where considerable plutonium residues resulting from safety tests were present [96].

Bikini Atoll Test Site

Nuclear tests were conducted at Bikini Atoll from 1946 to 1958. The regional fallout from the Bravo test caused widespread contamination of the Bikini Atoll and forced the evacuation of Marshallese people living on

Rongelap and Utrik Atolls [96]. About 50% of the fission yield associated with near-surface nuclear detonations was deposited on a local or regional scale [99].

The 167 inhabitants of Bikini Island were evacuated prior to the first nuclear test in 1946. In August 1968, resettlement of the Bikinian people on the atoll was allowed and about 100 people returned to the atoll. Because of concerns expressed by Bikini residents about the safety of living on the atoll, in 1975–1978 a set of additional investigations were carried out gathering data for more precise dose estimates [100]. In 1978, a tenfold increase in the body content of ^{137}Cs compared to those measured in 1970 was identified, based on whole body measurements [101] which were associated with consumption of coconut fluid due to the limited availability of drinking water. In August and September 1978, in response to the uptake of cesium in the population, the Bikinians who had returned to Bikini Atoll were relocated once again [96].

The total annual dose to the potential residents of the atoll (above the natural background dose) was estimated in 1999 to be 4.0–15 mSv, depending on the ratio of local to imported foods in the diets of inhabitants. Cs-137 was identified as the main radionuclide contributing to the dose, while contributions of ^{90}Sr , $^{239+240}\text{Pu}$, and ^{241}Am were much lower. The intake of marine foods, stored rain water and groundwater, and inhalation of resuspended soil together accounted for less than 1% of the dose [102].

Although full-scale remedial actions at the site have not been applied, a few remedial options for Bikini were implemented and studied in detail. The interventions were based on projected dose to a hypothetical critical group. The assessed dose of 15 mSv [103] was higher than the action dose level for intervention of 10 mSv accepted at that time [29]. Therefore, the implementation of remediation was considered to be highly justified. The key remedial options used were (1) washing-out ^{137}Cs from the soil with large quantities of salt water, (2) application of ameliorants such as zeolites to soil to make ^{137}Cs less available for plant uptake, (3) planting vegetation with high biomass to remove ^{137}Cs from the soil and then burial of the contaminated biomass, (4) soil removal (40 cm top layer), and (5) treating soil with potassium fertilizers [96].

Soil removal and treatment with potassium fertilizers were identified as the most effective measures.

However, soil removal would necessitate moving about 30,000 mature coconuts, pandanus and papaya trees, and breadfruit and was therefore a complex and extremely expensive operation. In contrast, soil removal from dwelling areas would be more cost effective and the best option was to remove soil from around each house and replace it with a layer of crushed coral to minimize external exposure and possible ingestion of the remaining soil. Consideration was also given to the potential ecological effects of the proposed potassium fertilizer treatment; at the planned levels there was no significant alteration expected in the soil chemistry and the potential transfer of potassium to groundwater was very low [104]. Application of potassium fertilizer would be needed every 4 or 5 years to maintain the low ^{137}Cs activity concentrations in local foods. It has been estimated that after the potassium treatment of soil and the replacement of soil from around the living areas, the dose rate based on the assumption that residents were consuming only local food would be reduced by as much as a factor of 10 or slightly higher compared with the levels before remedial actions application [105].

Maralinga Test Site

Tests of the UK nuclear program at the Maralinga test site located on the Nullarbor Plain in South Australia began in 1955. The site was abandoned in 1967 following the advent of the Nuclear Non-proliferation Treaty and the banning of atmospheric nuclear tests because the site was not suitable for underground tests. Seven nuclear tests were performed at the site, with approximate yields ranging from 1–27 kt. The site was also used for hundreds of minor trials from 1955 to May 1963, many of which were safety tests intended to investigate the effects of fire or nonnuclear explosions on atomic weapons [106].

Since 1967, numerous surveys have been carried out to characterize the contamination within the site. Substantial contamination densities were found at many test sites, particularly at Taranaki site where the Vixen B nuclear test was performed. Several square kilometers of land were contaminated to levels exceeding 300 kBq m^2 of ^{239}Pu . Some limited areas at the site were 10- or 100-fold more contaminated. Initially, plutonium contamination consisted of the following isotopes: ^{238}Pu ,

^{239}Pu , ^{240}Pu , and ^{241}Pu ; radioactive decay resulted in ^{241}Pu being gradually replaced with ^{241}Am [96].

The first efforts to remediate the Maralinga site began in 1964. Initially, contaminated debris and most of the remaining infrastructure were buried in pits at a depth of 2–3 m. The Pu content of each of these pits was not well quantified, but was estimated to be in kilograms quantities [106]. In 1967, further steps were taken to reduce the plutonium surface concentrations by turning over and mixing the surface soil. Following the remediation the site was closed with a series of fences erected to enclose the burial pits containing significant quantities of plutonium at Taranaki and other sites [96].

Later assessments made for nomadic Aborigines living mainly an outdoor lifestyle identified the inhalation dose pathway as the contributor to the total doses to both adults and children [107]. As a result of this survey, a much more extensive remediation project was initiated at the site, within which three sets of criteria were established for making decisions on the remediation process, namely:

1. Contaminated soil was removed where the levels of dispersed ^{241}Am exceeded 40 kBq m^{-2} averaged over 1 ha, or where contaminated particles exceeding 100 kBq were found, or where the density of particles exceeding 20 kBq was greater than 1 in 10 m^2 (Soil-Removal Criteria).
2. Once soil was removed the residual levels of dispersed contamination in the cleared area was not to exceed 3 kBq m^{-2} for ^{241}Am , averaged over 1 ha, and particulate contamination was to meet the Soil-Removal Criteria (Clearance Criteria).
3. Permanent occupancy and unrestricted land-use was only to occur where levels of dispersed contamination were less than 3 kBq km^{-2} for ^{241}Am , averaged over 3 km^2 , and the particulate contamination met the Soil-Removal Criteria (Unrestricted Use Criteria) [108–110].

To reduce the inhalation risk, the remediation criteria were guided by conservative principles and estimation of doses by realistic scenarios. These included the possibility of an Aboriginal group living for an entire year on the edge of the nonresidential area in regions of the highest activity permitted (approximately 3 kBq m^{-2} of ^{241}Am).

The first stage of the remediation project consisted of defining the boundaries at the sites contaminated with Pu (Taranaki, TMs, and Wewak), followed by bulk removal of contaminated soil from the three sites and burial in excavated burial trenches under at least 5 m of clean rock and soil. At the end of the remediation process, a compliance assessment was carried out to ensure that the whole Maralinga area had been made safe following the work undertaken during the 1994–2000 Maralinga project [96].

Following the remediation by removal and burial at depth of contaminated surface soil, all areas at Maralinga were shown to be within acceptable absorbed dose limits for all envisaged land uses. The restriction on permanent occupancy within the “restricted land-use” (nonresidential) boundary surrounding Taranaki was set as a precautionary measure as the inhalation doses for permanent occupancy of all but a few areas (essentially within the untreated plumes) were well below the 1 mSv a^{-1} limit for members of the public [96].

Remediation of Sites within Decommissioning of Facilities

The use of the term decommissioning implies that no further use of the facility (or part thereof) for its existing purpose is foreseen [111]. Decommissioning actions are taken at the end of the operating lifetime of a facility to remove it from service with due regard for the health and safety of workers and members of the public and the protection of the environment. Subject to national legal and regulatory requirements, a facility or its remaining parts may also be considered decommissioned if it is incorporated into a new or existing facility, or even if the site on which it is located is still under regulatory control or institutional control [112].

The actions taken during decommissioning will need to be such as to ensure the long protection of the public and the environment, and typically include reducing the levels of residual radionuclides in the materials and the site of the facility so that the materials can be safely recycled, reused, or disposed of as exempt waste or as radioactive waste, and the site can be released for unrestricted use or otherwise reused [111].

Remediation can be applied to both nuclear and nonnuclear contaminants that can affect human health

and the environment. Remediation encompasses *inter alia* site characterization, identification of remedial action alternatives, implementation of a remedial action, and ongoing monitoring to assure the confinement or containment of residual contamination.

There are many facilities with different characteristics around the world contaminated with radionuclides and therefore, according to the legislation, require decommissioning. They can include nuclear power reactors, radiochemical processing facilities, uranium and thorium mines, research institutions and hospitals, and many other human activities [28]. Therefore, remediation activities can be also very different and vary from decontamination of selected pieces of equipment to large-scale soil removal near some uranium-mining facilities or demolition of large concrete structures and placing them into final or interim waste storage [111].

In contrast, nuclear laboratories or institutions with a good standard of radiation protection can be decommissioned after limited remediation efforts although in all cases, the decommissioning process must be well planned and sufficient resources must be available. Over time, without proper arrangements being made for decommissioning, shut down facilities deteriorate and ultimately constitute a radiological hazard in their vicinity from direct external exposure to radiation to the public and potential release of radioactive material to the environment [112]. Currently, a substantial discharge of radionuclides has only been observed within a few decommissioning projects, largely at uranium-mining and uranium-milling facilities [113].

Remediation goals are usually based on dose constraints and derived action levels expressed in concentrations of radionuclides in various media. Because of the high cost, remediation is more often aimed at achieving “restricted reuse” or “fit-for-purpose reuse” as opposed to “unrestricted reuse,” though the latter would be the ideal goal. The remedial objectives for any restricted reuse are then based on meeting appropriate dose criteria associated with the intended land reuse. Where these doses are within the acceptable limit, regardless of the future uses of the site, then this enables a site to be released from any controls [61]. However, material originating from a released site needs to comply with the national requirements for

radiation protection for materials outside the scope of regulatory control [114].

The above aspects clearly show the importance of implementing a forward-looking planning process. While in the past, decommissioning and remediation activities have often been carried out as independent activities, optimization of effort, cost, impact, and risk can be achieved by carrying out decommissioning and remediation activities in an integrated manner as described in Falck et al. [112].

The high public concern regarding nuclear safety issues and the intention to build confidence to the remediation success has led to an acknowledgement of the need for stakeholder involvement. A system of the activities intended for providing transparency in remediation can be summarized as follows:

1. Development of an effective public-communication project policy (openness, transparency, truthfulness, effectiveness, timeliness etc.)
2. Development of the stakeholder involvement goals (no surprises, confidence generating)
3. Identification of relevant stakeholders to be involved into the decision conference
4. Selection of communication, consultation, and participation methods for training of project staff
5. Monitoring and review of success through local attitude survey; feedback from the stakeholders of different levels

As noted earlier, there are many differing facilities with distinct remediation goals, required remedial actions and further planned use of the contaminated site. Decommissioning of the Wismut uranium mine in Germany, based mainly on overview of remediation activities within decommissioning is an example of successful remedial activities typical for the decommissioning process.

Uranium mining in Saxony and Thuringia, known as Soviet-German Company (SDAG) Wismut, was in operation for more than 40 years. During that time, the Wismut company had mined around 231,000 t of uranium. Due to poor disposal practices in the first years of the operation, and routine natural radionuclides releases thereafter, vast areas mainly within the territory of the facility were contaminated. In total, the area of the Wismut Company covered 36.9 km², including 24.4 km² of mine dumps, 7.2 km² of tailings ponds, and

1.6 km² of open pit mine. The contamination of the site was represented by different radioactive materials, including: (1) scattered ore, (2) scattered tailings material (specific activity around 1–10 Bqg⁻¹ for ²²⁶Ra as the key nuclide, not in radioactive equilibrium), (3) lost uranium concentrate, (4) naturally leached uranium, for instance at sites from which waste dumps were removed [112].

In 1991, following integration of West and East Germany when the Federal Republic of Germany took responsibility for the SDAG Wismut Company, and the corresponding nuclear legacy, the Wismut Remediation Project was started. The project involves remedial activities at many sites, including four underground mines and two processing plants. In 1991, for a reuse of the sites, buildings, and materials contaminated by uranium mining activities, the German Commission on Radiological Protection recommended a reference level of effective dose of 1 mSv a⁻¹.

The extent of remedial measures which were undertaken at the site was justified based on site evaluation and remediation feasibility studies, rather than by applying uniform standards. Remediation of the Wismut legacy was viewed as an opportunity to return the affected land to productive use, thus enhancing the revitalization of the former mining areas in East Germany [112].

The Wismut site remediation took advantage of the open space available in unflooded mines and other sites to dispose of contaminated material in situ. The places used for disposal of contaminated material included:

- Underground storage sites (“dry” mine galleries above the groundwater level)
- Waste rock dumps (construction of cassettes, placement of the waste, coverage)
- Beach areas of tailings ponds (same technology as for dumps)
- Engineered hazardous waste disposal facility (constructed on top of a waste rock pile)

Special attention was given to the operational waste that resulted from the application of modern remediation technologies for the treatment of contaminated material. The contamination was concentrated in residues, reaching specific activities of up to 1,000 Bqg⁻¹ ¹²³⁸U and 20 Bqg⁻¹ ¹²²⁶Ra. Special techniques for the immobilization of these residues needed to be applied. Despite the lack of the space available for

disposal, separation of waste according to type and level of contamination and further recycling of low contaminated or uncontaminated materials were the key elements of the waste management practice at the Wismut site. In particular, salvaged uncontaminated metallic scraps and scraps with a total surface alpha activity lower than 0.5 Bqcm^{-2} have been sold for smelting [112].

Former waste rock piles cannot always be completely remediated and they can be released for restricted reuse. In particular, such areas could be used for planting new forests, or in some cases even for fodder production or various industrial or trade applications. Forestry was an attractive cost-effective option which was considered sustainable in the long term.

Considerable efforts were made over the past years for the benefit and protection of the population inhibiting areas affected by mining activities at the Wismut site. The significant reduction of adverse environmental impacts and the reutilization of these areas for industrial use, housing construction, or as a nature reserve have revitalized these regions. The “New Horizons through Remediation” program became a key element of Remediation. Integrated state-of-the-art technologies from around the world were applied to conduct decommissioning and remediation of the radioactive legacies left by the uranium mining in an efficient and cost-effective way leading to sustainable solutions. Remediation has had a benefit for regional development at two former mining towns, leading to rebirth of the health spa of Schlema (Fig. 10) and the development of the Federal Garden and Landscape Exhibition in 2007 (BUGA, 2007) to be hosted by the towns of Ronneburg and Gera [112].

Overall, the Wismut Project provides evidence of the potential success of a value-added approach with no high additional (or not reimbursable) costs, if remediation is efficiently undertaken with a well-defined reutilization in mind. Such successful utilization goals can only be developed in cooperation with the envisaged future user (i.e., municipality or developer), the community and the regulatory authorities. If consensus is achieved with the stakeholders prior to remedial actions, then “remediation by objectives” becomes practicable, i.e., the objectives for the individual remedial steps can be determined, consistent with the ultimate utilization goal [112].



Environmental Countermeasures and Restoration.

Figure 10

Former mining area near Schlema before (1991) and after (2009) remediation (<http://www.wismut.de/>)

Decision-Aiding Tools

The efficiency of countermeasures or remediation depends on many factors: the specific contamination scenario, the properties of the contaminants involved, characteristics of the affected environment, the perception of countermeasures or remedial actions by the local population amongst many others. The effectiveness of such actions, cost, and resources required for their implementation vary considerably with these various factors; this in turn affects the suitability of different countermeasures or remedial options for a given situation. Therefore, generalized recommendations which do not account for the diversity of local site-specific conditions can result in inadequate decision-making and may not be feasible to execute.

These multiple factors must be reconciled within a single optimization approach and there is a significant challenge in finding a balanced solution to the multidimensional problem each contamination scenario presents. This task can be facilitated with environmental decision support systems (EDSS), capable of providing advice on countermeasure/remediation strategies at different levels of the decision-making process, which take into account temporal and spatial variation in each factor. A conceptual framework and basic elements of such systems (EDSS) intended for support in remediation of contaminated areas is given in Fig. 11.

Following the Chernobyl accident, various EDSS, or models which underpin them, have been produced, such as FORCON [115], SAVE [40], RESTORE [116, 117], and CESAR [118] which have enhanced the ability to optimize remediation strategies in contaminated areas.

Decision-aiding techniques are useful tools to determine the level and kind of protection which will provide the best solution given the various factors and constraints involved, taking into account inherent uncertainties and value judgments. This can only be achieved on the basis of a clear identification of

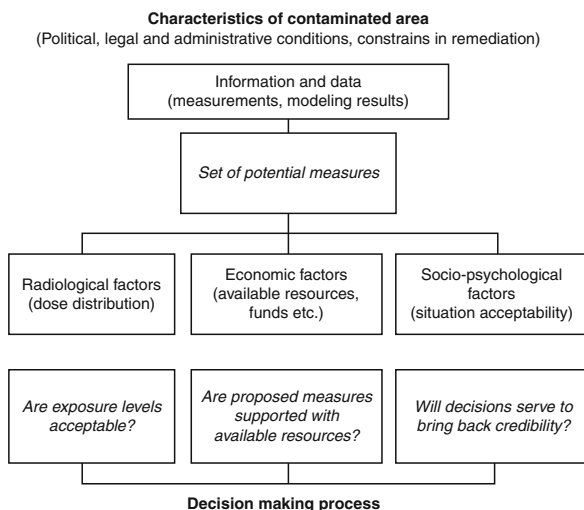
available alternative management options and the factors and constraints involved in the processes. Site-specific quantification is a difficult and time-consuming step in the decision-making process because all necessary data must be gathered or generated by models to then use in combination with decision-aiding techniques to optimize the decision-making process [119]. Decision-aiding techniques have been widely used to select preferable remedial options for contaminated areas, starting in the early 1980s when ICRP recommended the use of cost-benefit analyses [34, 35]. In these analyses, the cost of environmental deterioration as a result of radiological contamination which could be averted using protective measures was compared with the cost of the protective measures [34, 35, 120, 121]. For decisions of greater complexity, including considerations of large-scale remedial, decision-making can be improved by complementing the above methods with multi-criteria analysis [13].

Cost-effectiveness and cost-benefit techniques are based on balancing the costs of management option implementation against the expected reduction of the collective dose, which is expressed in monetary value. The definition and use of the monetary value of collective doses, introduced by ICRP in Publication 22 [120], has been a key element of this approach; the implementation of this approach has been, and currently is, the subject of many debates on ethical grounds.

There are some clear differences between these two techniques. Cost-effectiveness analysis is intended mainly to eliminate low cost-effective management options and to rank the remaining options (see Fig. 12). The main elements of the approach are as follows:

1. Characterization of management options in terms of their cost and the corresponding residual collective doses following their implementation.
2. Selection of cost-effective options for which there are no alternative options either to reach the same residual collective dose at a lower protection cost or for the same protection cost to reach a lower residual collective dose.

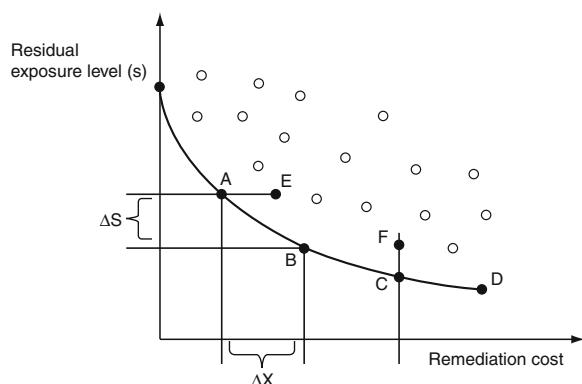
This process is illustrated graphically in Fig. 12. Each management option is represented by a dot and all those dots situated along the cost-effectiveness curve



Environmental Countermeasures and Restoration.

Figure 11

Key elements of the environmental decision support system (EDSS) for remedial measures on contaminated areas (Adopted from Shershakov et al. [13])



Environmental Countermeasures and Restoration.

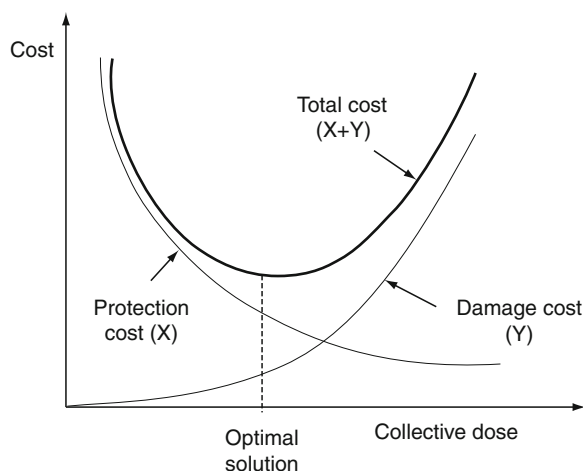
Figure 12

Cost-effectiveness analysis (Adopted from Lochard and Schneider [122])

represent cost-effective options. For example option (A) allows for an equivalent residual collective at a lower cost than (E), and option (C) gives a lower residual collective dose for the same cost as (F). All the options which are above on the curve are called “totally dominated” (worse) options and must be discarded from further consideration (Lochard and Schneider 2001 [122]).

Cost-effectiveness is intended to assess the “marginal cost” of every management option, which has to be compared against the closest less or more expensive options. If a small additional cost would lead to a much higher effectiveness in terms of risk reduction, the option is not cost effective. Finally, each cost-effective option can be characterized by the increase in cost from one option to the next (DX) and the corresponding decrease in collective dose (DS). The quotient (DX/DS) is called the cost-effectiveness ratio and provides a basis for ranking the various options. However, determination of the cost-effectiveness curve and the corresponding cost-effectiveness ratios does not provide a basis for the selection of the optimum option. This is achieved by the introduction of a reference value for the cost-effectiveness ratio in the framework of the cost-benefit analysis [122].

The most widespread approach to performing the second technique, cost-benefit analyses is based on the determination, for each management option,



Environmental Countermeasures and Restoration.

Figure 13

Cost-benefit analysis applied for the identification and justification of the optimal management option [34]

of the sum of remediation costs and the cost of residual collective dose (expressed in monetary terms) resulting from that strategy, and then comparing them to identify which option has the lowest total cost. This option can be accepted as the optimum solution, as illustrated in Fig. 13 [34].

There are many factors to be taken into account as part of the process of remediation justification and some of them cannot be quantified either in monetary terms or in terms of any resources, although they may be quantifiable in other units or ranking parameters. In such a case the most straightforward way to justify an optimized remediation strategy is the application of multi-attribute-utility analysis (MAUA). This technique was first recommended by ICRP for the optimization of radiation protection efforts in ICRP Publication 55 [35]. The most recent ICRP Publications 103 and 104 emphasize the great impact of social and political factors on decisions regarding remediation which also encourage the use of MAUA as one of the basic approaches for remediation optimization [10, 31].

The basic principle of the multi-attribute-utility analysis technique is to construct a scoring scheme (or multi-attribute-utility function) for each option on the basis of all significant criteria which characterize

the situation (i.e., cost of remediation, collective or individual doses, perception of management options by the population, etc.). When the different remedial options are identified, the MAUA defines the relevant criteria for the decision process. The criteria are specified in compliance with specific strategies to ensure remediation success. These criteria normally include: economic attributes, human health impact and public safety attributes, socio-psychological attributes, environmental secondary effects and ethical considerations, criteria based on public opinions, and the interests of various population groups.

Every remediation alternative must be evaluated according to some or all of the above criteria, either quantitatively or qualitatively. Because of their diversity, possible options will be ranked differently for each criterion. In further steps, weighting-coefficients are assigned to each criterion to account for their relative importance. This is the most important and often difficult step in MAUA. Nevertheless, several techniques are available to derive this set of values, and whatever method is used, the choice of weighting coefficients should be justified. Formally, each option is qualified by its total utility (U) calculated as follows:

$$U_i = \sum k_j \cdot u_{ij},$$

Where i is the remediation alternative number; j is the criterion number; k_j is the weighting coefficient showing the relative importance of each option (normalized $\sum k_j = 1$); u_i is the utility of measure i .

The single utility associated with each criterion can be defined either as a linear function of the value expressing the criteria or as a nonlinear function to incorporate the preferences of the decision makers into the analysis. For example, it is possible to define utility functions which incorporate risk aversion according to the level of individual or collective exposures [35].

Finally, the management options which lead to the highest total utility should be selected. Since most of the weighting factors generally rely on decision makers' judgments, it is highly recommended that a sensitivity analysis be performed using different sets of weighting coefficients to test the "robustness" of the results.

Based on such an approach, a few specific decision-support systems for remediation of different

environments: aquatic, forest, and rural were suggested and used for the long-term optimization of remediation strategies following the Chernobyl accident [85, 123–127].

Future Directions

Different nuclear applications have contributed to the rapid development of science and technology within the second half of the twentieth century. The use of numerous nuclear activities has, in many countries around the world, led to the contamination of significant parts of some ecosystems, which require remediation. Due to the variety of sources and the characteristics of the receiving environment, the scenario of contaminated sites varies greatly. Nuclear weapon testing and severe radiation accidents, in particular, have also greatly contributed to global concern about the need for reliable and comprehensive radiation risk assessment. In response, the development of countermeasures and remediation constitute two major radiation protection tools that can alleviate the radiological consequences of emergencies and reducing the radiation exposure from existing contamination, respectively.

Currently, the quality of assessments of the consequences of the contamination of the environment varies depending on the amount of available data on the environmental behavior of different radionuclides in the environment. While the data available on some key dose forming radionuclides, especially those released on the major accidents, are reasonably good, that of many other radionuclides is rather patchy. Therefore, there are some notable deficiencies in information on environmental behavior of many important radionuclides [128] and also in the associated radiation effects on many biological and ecological end points particularly in plants and animals. In some cases, contaminated sites already released for general use have had to be reclassified because of an insufficient understanding or quantification of the radiation risk at the time when they were released. The need for remediation of such sites was not fully appreciated as soon as it might have been if better information had been available on the environmental pathways of exposure.

Appropriate response to contamination of the environment with radionuclides requires a comprehensive multidisciplinary approach as outlined in this chapter. The paper is focused on actions to prevent exposure of humans but the challenge is greater than this in that the impact of countermeasures or remediation on wildlife is difficult to quantify on the basis of current knowledge. Therefore, there is a significant future challenge to consider the need for, and potential implementation of, a holistic framework for the response to environmental contamination considering both the protection of human beings and the environment itself.

Bibliography

Primary Literature

1. IAEA (2007) IAEA safety glossary: terminology used in nuclear, radiation, radioactive waste and transport safety. International Atomic Energy Agency, Vienna
2. Eisenbud M, Gessel T (1997) Environmental radioactivity from natural, industrial and military sources, 4th edn. Academic Press, San Diego, CA (ISBN 0-12-235154-1)
3. IAEA (1987) Dosimetric and medical aspects of the radiological accident in Goiania in, 1987. IAEA TECDOC-1009. International Atomic Energy Agency, Vienna
4. IAEA (1999) Technologies for the remediation of radioactively contaminated sites, International Atomic Energy Agency IAEA-TECDOC-1086. International Atomic Energy Agency, Vienna
5. UNSCEAR (2000) United Nations Scientific Committee on the effects of atomic radiation. Report to the general assembly. sources and effects of ionizing radiation. Scientific Annex C, UNSCEAR, New York, pp 157–252
6. Aarkrog A (2001) Manmade radioactivity. In: Van der Stricht E, Kirchman R (eds) Radioecology: radioactivity and ecosystems. Fortemps, Liege, pp 55–78
7. IAEA (2002) Radiation legacy of the 20th century: environmental restoration IAEA-TECDOC-1280. International Atomic Energy Agency, Vienna
8. IAEA (2003) Remediation of areas contaminated by past activities and accidents safety requirements IAEA safety standards series No. WS-R-3. International Atomic Energy Agency, Vienna
9. González AJ (2009) International approaches to remediation of territorial radioactive contamination. In: Voigt G, Fesenko S (eds) Remediation of contaminated environments. Elsevier, Amsterdam, pp 1–42
10. ICRP (2008) Scope of radiological protection control measures ICRP publication 104 Annals ICRP 37(5). Elsevier, Amsterdam
11. Voigt G, Eged K, Howard BJ, Kis Z, Nisbet AF, Oughton DH, Rafferty B, Salt CA, Smith JT, Vandenhove H (2000) A wider perspective on the selection of countermeasures. Rad Prot Dosim 92:45–48
12. Fesenko SV, Alexakhin RM, Balonov MI, Bogdevitch IM, Howard BJ, Kashparov VA, Sanzharova NI, Panov AV, Voigt G, Zhuchenka YM (2007) An extended critical review of twenty years of countermeasures used in agriculture after the Chernobyl accident. Sci Total Environ 383:1–24
13. Shershakov V, Fesenko S, Kryshev I, Semioshina N (2009) Decision-aiding tools for remediation strategies. In: Voigt G, Fesenko S (eds) Remediation of contaminated environments. Amsterdam, Elsevier, pp 123–178
14. Jacob P, Fesenko SV, Firsakova SK, Likhtarev IA, Schotola C, Alexakhin RM, Zhuchenko YM, Kovgan L, Sanzharova NI, Ageyets V (2001) Remediation strategies for rural territories contaminated by the Chernobyl accident. J Environ Radioactiv 56:51–76
15. Lochard J (2004) Living in contaminated territories: a lesson in stakeholder involvement. In: Current trends in radiation protection (EDP Sciences), pp 211–220
16. Fesenko SV, Alexakhin RM, Balonov MI, Bogdevich IM, Howard BJ, Kashparov VA, Sanzharova NI, Panov AV, Voigt G, Zhuchenko YuM (2006) Twenty years' application of agricultural countermeasures following the Chernobyl accident: lessons learned. J Radiol Prot 26(4):1–9
17. Fesenko S, Zeiller L, Voigt G (2009) Site characterisation and measurement strategies for remediation purposes. In: Voigt G, Fesenko S (eds) Remediation of contaminated environments. Elsevier, Amsterdam, pp 43–122
18. Shaw G (2007) Introduction. In: Shaw G (ed) Radioactivity in the terrestrial environment, Radioactivity in the environment. Elsevier, Amsterdam, pp 1–19
19. Iranzo E, Richmond CR (1987) Plutonium contamination twenty years after the nuclear accident in Spain. Oak Ridge National Laboratory, Tennessee
20. Wrenn MMN (1974) Environmental levels of plutonium and transplutonium elements WASH 1359 (Atomic Energy Commission). Washington, DC, USA, p 89
21. Eriksson M (2002) On weapons plutonium in the arctic environment (Thule, Greenland) Riso-R-1321(EN). Riso National Laboratory, Roskilde, Denmark
22. Strand P, Balonov M, Bewers M, Howard BJ, Tsaturov YS, Salo A, Aarkrog A (eds) (1997) Arctic pollution issues. radioactive contamination. A report from an international expert group to the arctic monitoring and assessment programme. Oslo, Norwegian Radiation Protection Authority, p 160
23. Ramzaev V, Mishin A, Golikov V, Argunova T, Ushnitsi V, Zhuravskaya A, Sobakin P, Brown J, Strand P (2008) Radioecological studies at the Kraton-3 underground nuclear explosion site in 1978–2007: a review In: Strand P, Brown J, Jølle T (eds) Proceedings of the international conference on radioecology and environmental radioactivity, NRPA, Norway, pp 448–451
24. Alexakhin RM, Fesenko SV, Sanzharova NI (1996) Serious radiation accidents and the radiological impact on agriculture. Radiat Prot Dosim 64:37–42
25. Alexakhin RM, Buldakov LA, Gubanov VA, YeG D, Ilyin LA, Kryshev II, Linge II, Romanov GN, Savkin MN, Saurov MM,

- Tikhomirov FA, Kholina YuB (2004) Large radiation accidents: consequences and protective countermeasures. Moscow, IzdAT Publisher
26. Alexakhin RM (2009) Remediation of areas contaminated after radiation accidents. In: Voigt G, Fesenko S (eds) Remediation of contaminated environments. Elsevier, Amsterdam, pp 179–224
 27. Cancio DJ, Gutierrez C, Rujz AS (1993) Radiological considerations related with the restoration of a phosphogypsum disposal site in Spain, European Commission Doc. XI-5027/94. In: Proceedings international symposium remediation and restoration of radioactive-contaminated sites in Europe, Antwerp. pp 645–652
 28. IAEA (1999) Compliance monitoring for remediated sites, international atomic energy agency, IAEA-TECDOC-1118. International Atomic Energy Agency, Vienna
 29. IAEA (1996) International basic safety standards for protection against ionizing radiation and for the safety of radiation sources. IAEA safety series No.115. International Atomic Energy Agency, Vienna
 30. ICRP (1991) Recommendations of the international commission on radiological protection ICRP publication 60 Annals ICRP 21(1–3). Elsevier, Amsterdam
 31. ICRP (2007) The 2007 recommendations of the international commission on radiological protection ICRP Publication 103 Annals ICRP 37(2–4). Elsevier, Amsterdam
 32. ICRP (2006) Assessing dose of the representative person for the purpose of radiation protection of the public and the optimisation of radiological protection: broadening the process publication 101 Annals ICRP 36(3). Elsevier, Amsterdam
 33. ICRP (1999) Protection of the public in situations of prolonged radiation exposure ICRP Publication 82 Annals ICRP 29(1/2). Elsevier, Amsterdam
 34. ICRP (1983) Cost-benefit analysis in the optimisation of radiation protection ICRP publication 37. Annals ICRP 10(2/3). Pergamon Press, Oxford
 35. ICRP (1989) Optimisation and decision-making in radiological protection. ICRP publication 55 Annals ICRP 20 (1). Pergamon Press, Oxford
 36. IAEA (2002) Non-technical factors impacting on the decision making processes in environmental remediation IAEA-TECDOC-1279. International Atomic Energy Agency, Vienna
 37. IAEA (2000) Site characterisation techniques used in environmental restoration activities, International Atomic Energy Agency IAEA-TECDOC-1148. International Atomic Energy Agency, Vienna
 38. Fesenko SV, Colgan A, Sanzharova NI, Lissianski KB, Vazquez C, Guardans R (1997) The dynamics of the transfer of caesium-137 to animal fodder in areas of Russia affected by the Chernobyl accident and doses resulting from the consumption of milk and milk products. Radiat Prot Dosim 69:289–299
 39. Fesenko SV, Jacob P, Alexakhin R, Sanzharova NI, Panov A, Fesenko G, Cecille L (2001) Important factors governing exposure of the population and countermeasure application in rural settlements of the Russian Federation in the long term after the Chernobyl accident. J Environ Radioactiv 56:77–98
 40. Howard BJ, Wright SM (eds) (1999) Spatial analysis of vulnerable ecosystems in Europe: spatial and dynamic prediction of radiocaesium fluxes into European foods (SAVE). Final report. Commission of European Communities, Brussels
 41. Howard BJ, Liland A, Beresford NA, Andersson KG, Cox G, Gil JM, Hunt J, Nisbet A, Oughton D, Voigt G (2004) A critical evaluation of the strategy project. Radiat Prot Dosim 109:63–67
 42. Sanzharova NI, Fesenko SV, Alexakhin RM, Anisimov VS, Kuznetsov VK, Chernyayeva LG (1994) Changes in the forms of ^{137}Cs and its availability in plants as dependent on properties of fallout after the Chernobyl Nuclear Power Plant accident. Sci Total Environ 154:9–22
 43. Dale P, Robertson I, Toner M (2008) Review: radioactive particles in dose assessments. J Environ Radioactiv 99:1589–1595
 44. Alexakhin RA, Korneev NA (eds) (1991) Agricultural Radioecology. Ecology, Moscow (in Russian)
 45. Prister BS, Perepelyatnikov GP, Perepelyatnikova LV (1993) Countermeasures used in the Ukraine to produce forage and animal food products with radionuclide levels below intervention limits after the Chernobyl accident. Sci Total Environ 137:183–198
 46. Howard BJ (2000) The concept of radioecological sensitivity. Radiat Prot Dosim 92:29–34
 47. Alvarez-Farizo B, Gil JM, Howard BJ (2009) Impacts from restoration strategies: assessment through valuation workshops. Ecol Econ 68:287–297
 48. Howard BJ, Beresford NA, Nisbet A, Cox G, Oughton DH, Hunt J, Alvarez B, Andersson KG, Liland A, Voigt G (2005) The STRATEGY project: decision tools to aid sustainable restoration and long-term management of contaminated agricultural ecosystems. J Environ Radioactiv 83:275–295
 49. Alexakhin RM (1993) Countermeasures in agricultural production as an effective means of mitigating the radiological consequences of the Chernobyl accident. Sci Total Environ 137:9–20
 50. Prister B, Loshchilov N, Perepelyatnikova L, Perepelyatnikov G, Bondar P (1992) Efficiency of measures aimed at decreasing the contamination of agricultural products in areas contaminated by the Chernobyl NPP accident. Sci Total Environ 112:9–87
 51. Howard BJ, Beresford NA, Voigt G (2001) Countermeasures for animal products: a review of effectiveness and potential usefulness after an accident. J Environ Radioactiv 56:115–137
 52. Bogdevitch I, Sanzharova N, Prister B, Tarasiuk S (2002) Countermeasures on natural and agricultural areas after Chernobyl accident. In: Kolejka J (ed) Role of GIS in lifting the cloud off Chernobyl, Kluwer, Dordrecht, pp 147–158
 53. Junker H, Jensen JM, Jørgensen HB, Roed J, Andersson KG, Kofman PD, Bøllehuus E, Baxter L, Grebenkov A (1998) Chernobyl bioenergy project, ELSAMPROJEKT report no. EP 8204-98-n600hju
 54. Andersson K, Roos P (2009) Effectiveness of remedial techniques for environments contaminated by artificial radionuclides. In: Voigt G, Fesenko S (eds) Remediation of contaminated environments. Elsevier, Amsterdam, pp 375–426

55. Andersson KG, Roed J, Eged K, Kis Z, Voigt G, Meckbach R, Oughton DH, Hunt J, Lee R, Beresford NA, Sandalls FJ (2003) Physical countermeasures to sustain acceptable living and working conditions in radioactively contaminated residential areas, RISOE-R-1396(EN), ISBN 87-550-3190-0
56. Andersson KG, Brown J, Mortimer K, Jones JA, Charnock T, Thykier-Nielsen S, Kaiser JC, Proehl G, Nielsen SP (2008) New developments to support decision-making in contaminated inhabited areas following incidents involving a release of radioactivity to the environment. *J Environ Radioactiv* 99:439–454
57. Brown J, Mortimer K, Andersson K, Duranova T, Mrskova A, Hänninen R, Ikäheimonen T, Kirchner G, Bertsch V, Gallay F, Reales N (2008) Generic handbook for assisting in the management of contaminated inhabited areas in Europe following a radiological emergency, parts I–V, Final report of the EC-EURANOS project activities CAT1RTD02 and CAT1RTD04, EURANOS(CAT1)-RP(07). Health Protection Agency, UK
58. Nisbet A, Brown J, Jones A, Rochford H, Hammond D, Cabianga T (2009) HPA-RPD-064 – UK recovery handbooks for radiation incidents, HPA, Didcot
59. Nisbet AF, Brown J, Howard BJ, Beresford NA, Ollagnon H et al (2010) Decision aiding handbooks for managing contaminated food production systems, drinking water and inhabited areas in Europe. *Radioprotection* 45(5):S23–S37
60. Salt CA, Rafferty B (2001) Assessing potential secondary effects of countermeasures in agricultural systems: a review. *J Environ Radioactiv* 56:99–114
61. IAEA (2006) Environmental consequences of the Chernobyl accident and their remediation: twenty years of experience report of the Chernobyl forum expert group environment. International Atomic Energy Agency, Vienna
62. Oughton D, Bay-Larsen I, Voigt G (2009) Environmental, social, ethical and economic aspects of remediation. In: Voigt G, Fesenko S (eds) Remediation of contaminated environments. Elsevier, Amsterdam, pp 427–452
63. Oughton D, Forsberg E-M, Bay I, Kaiser M, Howard B (2004) An ethical dimension to sustainable restoration and long-term management of contaminated areas. *J Environ Radioactiv* 74:171–183
64. Gould P (1990) Fire in the rain: the democratic consequences of Chernobyl. Polity Press, Cambridge
65. Heriard Dubreuil G, Lochard J, Girard P, Guyonnet JF, Le Cardinal G, Lepicard S, Livolsi P, Monroy M, Ollagnon H, Pena-Vega A, Pupin V, Rigby J, Rolevitch I, Schneider T (1999) Chernobyl post-accident management: the ETHOS project. *Health Phys* 77:361–372
66. Eged K, Kis Z, Voigt G, Andersson KG, Roed J, Varga K (2003) Guidelines for planning interventions against external exposure in industrial area after a nuclear accident – Part I: a holistic approach of countermeasure application, GSF-Bericht 01/03, GSF – Forschungszentrum für Umwelt und Gesundheit GmbH, Neuherberg, ISSN 0721-1694, 67 p
67. IAEA (1994) Handbook of parameter values for radionuclide transfer in the temperate environments. Technical Report Series 364, International Atomic Energy Agency, Vienna
68. Nisbet A, Howard B, Beresford N, Voigt G (2005) Stakeholder involvement in post-accidental management. *J Environ Radioactiv* 83:259–435
69. Beresford NA, Howard BJ, Barnett CL, Rantavaara A, Rissanen K, Reales N, Gallay F, Papachristodoulou C, Ioannides K, Nisbet AF, Hesketh N, Oughton D, Bay I (2005) Compendium of countermeasures for the management of food production systems. Deliverable TN(05)-01 (including CD) for EURANOS project (Cat1). Available at <http://www.euranos.fzk.de>
70. Nikipelov BV, Romanov GN, Buldakov LA, Babaev NS, Kholina YuB, Mikerin EI (1987) The radiation accident in the South Urals in 1957. *Atomnaya Energia* 67:74–80 (in Russian)
71. Burnazyan AI (1990) Results of the studies and experience of mitigation of consequences of the accidental contamination of the territory by uranium fission products. Energoatomizdat, Moscow (in Russian)
72. Sokolov VE, Krivolutsky DA (eds) (1993) Environmental impacts of radioactive contamination in the South Urals. Nauka, Moscow, 336 p ISBN 5-02-005719-3 (in Russian)
73. Akleev AV, Kiselev MF (2001) Ecological and medical impacts of the 1957 accident at the PP “Mayak”. RF Ministry of Health, Moscow
74. Romanov GN, Drozhko EG, Nikipelov BV, Teplyakov IG, Shilov VP (1993) Summing up: restoration of the economic activity. In: Sokolov VE, Krivolutsky DA (eds) Environmental impacts of radioactive contamination in the South Urals. Moscow, Nauka (in Russian)
75. Fetisov VI, Romanov GN, Drozhko EG (1993) Practice and problems of environment restoration at the location of the industrial association Mayak, 1994, European Commission Doc. XI-5027/94. In: Proceedings international symposium remediation and restoration of radioactive-contaminated sites in Europe, Antwerp, pp 507–521
76. Espinosa A, Aragon A, Stradling N, Hodgson A, Birchall A (1998) Assessment of doses to adult members of the public in Palamares from inhalation of plutonium and americium. *Radiat Prot Dosim* 79(1–4):161–164
77. Izrael YuA, Kvasnikova EV, Nazarov IM, Fridman ShD (1994) Global and regional contamination of the territory of the European part of the former USSR by ¹³⁷Cs. *Meteorol Hydrol* 5:5–9 (in Russian)
78. Izrael Yu (1999) Radioactive fallout after nuclear explosions and accidents. *Environ Monit Assess* 59(1):121–122
79. Sanzharova NI, Fesenko SV, Kotik VA, Spiridonov SI (1996) Behaviour of radionuclides in meadows and efficiency of countermeasures. *Radiat Prot Dosim* 64:43–48
80. Ratnikov AN, Vasiliev AV, Alexakhin RM, Krasnova EG, Pasternak AD, Howard BJ, Hove K, Strand P (1998) The use of hexacyanoferrates in different forms to reduce radiocesium contamination of animal products in Russia. *Sci Total Environ* 223:167–176
81. Strand P, Balonov M, Travnikova I, Hove K, Skuterud L, Prister B, Howard BJ (1999) Fluxes of radiocaesium in selected rural study sites in Russia and Ukraine. *Sci Total Environ* 231:159–171

82. Fesenko SV, Voigt G, Spiridonov SI, Sanzharova NI, Gontarenko IA, Belli M, Sansone U (2000) Analysis of the contribution of forest pathways to the radiation exposure of different population groups in Bryansk region of Russia. *Radiat Environ Biophys* 39:291–300
83. Shutov VN, Bruk GYa, Basalaeva LN, Vasiletskiy VN, Ivanova NP, Karlin IS (1996) The role of mushrooms and berries in the formation of internal exposure doses to the population of Russia after the Chernobyl accident. *Radiat Prot Dosim* 67:55–64
84. Tikhomirov FA, Shcheglov AI (1994) Main investigation results on the forest radioecology in the Kyshtym and Chernobyl accident zones. *Sci Total Environ* 157:45–57
85. Fesenko SV, Voigt G, Spiridonov SI, Gontarenko IA (2005) Decision making framework for application of forest countermeasures in the long term after the Chernobyl accident. *J Environ Radioactiv* 82:143–166
86. IAEA (2006) Radiological conditions in the Dnieper River Basin: assessment by an international expert team and recommendations for an action plan radioecological assessment report series. International Atomic Energy Agency, Vienna
87. Smith JT, Kudelsky AV, Ryabov IN, Haddingh RH, Bulgakov AA (2003) Application of potassium chloride to a Chernobyl contaminated lake: modelling the dynamics of radiocaesium in an aquatic ecosystem and decontamination of fish. *Sci Total Environ* 305:217–227
88. Kudelsky AV, Smith JT, Petrovich AA (2002) An experiment to test the addition of potassium to a non-draining lake as a countermeasure to ^{137}Cs accumulation in fish. *Radioprotection Colloques* 37:621–626
89. Roed J, Andersson KG, Barkovsky AN, Fogh CL, Mishine AS, Ponamarev AV, Ramzaev VP (2006) Reduction of external dose in a wet-contaminated housing area in the Bryansk Region. Russia *J Environ Radioactiv* 85:265–279
90. Rosenthal JJ, Cunha PO, Oliveira CA (1989) Aspects of the initial and recovery phases of the radiological accident in Goiania. In: *Proceedings international symposium recovery operations in the event of a nuclear accident or radiological emergency*, IAEA, Vienna, 593–597
91. Oliveira AR, Hunt JG, Valverde NJL, Brandao-Mello CE, Farina R (1991) Medical and related aspects of the Goiania accident: an overview. *Health Phys* 60(1):17–24
92. IAEA (1998) Dosimetric and medical aspects of the radiological accident in Goiania in, 1987. IAEA TECDOC-1009. International Atomic Energy Agency, Vienna
93. Rosenthal JJ, Almeida CE, Mendonca AH (1991) The radiological accident in Goiania, the remedial action. *Health Phys* 60:7–15
94. Anspaugh LR, Church BW (1986) Historical estimates of external γ exposure and collective external γ exposure from testing at the Nevada Test Site. I. Test series through Hardtack II, 1958. *Health Phys* 51:35–51
95. Donaldson LR, Seynour AH, Nevissi'k AE (1997) University of Washington's radioecological studies in the Marshall islands, 1946–1977. *Health Phys* 73:214–222
96. Danesi PR (2009) Remediation of sites contaminated by nuclear weapon tests in the Marshall Islands and Australia. In: Voigt G, Fesenko S (eds) *Remediation of contaminated environments*. Elsevier, Amsterdam, pp 225–262
97. Carlsen TM, Peterson LE, Ulsh BA, Werner CA, Purvis KL, Sharber AC (2001) Radionuclide contamination at Kazakhstan's semipalatinsk test site: implications on human and ecological health. Lawrence Livermore National Laboratory, Livermore, CA. UCRL-JC-143920
98. Bennet BG (2000) Worldwide panorama of radioactive residues in the environment. In: *Restoration of environments with radioactive residues*, Proceedings International Symposium Arlington, Virginia, USA, 29 Nov.–3 Dec. 1999, IAEA, Vienna, pp 11–24
99. Hamilton TF (2004) Linking legacies of the cold war to arrival of anthropogenic radionuclides in the oceans through the 20th century. In: Livingston HD (ed) *Radioactivity and the environment*, vol 6. Elsevier Science, Amsterdam, pp 30–87
100. Robison WL, Phillis WA, Colsher CS (1977) Dose assessment at Bikini Atoll, Rep. UCRL-51879, Part 5, Lawrence Livermore National Laboratory, Livermore
101. Miltenberger RP, Greenhouse NA, Lessard ET (1980) Whole body counting results from 1974 to 1979 for Bikini Island residents. *Health Phys* 39:395–507
102. Robinson WL, Noshkin VE (1999) Radionuclide characterization and associated dose from long-live radionuclides in close-in fallout delivered to the marine environment at Bikini and Enewetak Atolls. *Sci Total Environ* 237–238:311–327
103. Simon SL, Graham JC (1997) Findings of the first comprehensive radiological monitoring program of the Republic of the Marshall Islands. *Health Phys* 73:66–85
104. IAEA (1998) Characterization of radioactively contaminated sites for remediation purposes, IAEA-TECDOC-1017. International Atomic Energy Agency, Vienna
105. Robison WL, Bogen KT, Conrado CL (1997) An updated assessment for resettlement options at Bikini Atoll U.S. nuclear test site. *Health Phys* 73:100–114
106. Symonds JL (1985) A history of British atomic tests in Australia. Australian Government Publishing Service, Canberra
107. Haywood SM, Smith JG (1992) Assessment of potential doses at the Maralinga and Emu test sites. *Health Phys* 63:624–630
108. Parkinson A (2000) Maralinga rehabilitation project. National Conference-MAPW Australia, <http://www.mapw.org.au/conferences/mapw2000/papers/parkinson.html>
109. Williams GA (ed) (1990) Inhalation hazard assessment at Maralinga and Emu, Technical Report ARL/TR087
110. Williams GA, Cooper MB, Martin LJ (1998) Plutonium contamination at Maralinga – dosimetry and clean-up criteria. In: *Proceedings AIOH-98 17th annual conference of Australian institute of occupational hygienists (AIOH)*, December 1998, pp 117–126
111. IAEA (2005) Selection of decommissioning strategies: issues and factors. Report by an expert group IAEA-TECDOC-1478. International Atomic Energy Agency, Vienna

112. Falck WE, Seitz R, Pearl M, Audet M, Schmidt P, Fernandes HM (2009) Remediation as part of the decommissioning of nuclear facilities. In: Voigt G, Fesenko S (eds) Remediation of contaminated environments. Elsevier, Amsterdam, pp 263–374
113. NEA (2006) Releasing of the nuclear installations – a status report. NEA Publication No. 6187. OECD-Nuclear Energy Agency (OECD-NEA), Paris
114. IAEA (2004) Application of the concepts of exclusion, exemption and clearance, IAEA safety standard series No. RS-G-1.7. International Atomic Energy Agency, Vienna
115. Fesenko SV, Sanzharova NI, Wilkins BT, Nisbet AF (1996) FORCON: local decision support system for the provision of advice in agriculture – methodology and experience of practical implementation. Radiat Prot Dosim 64(1/2):157–164
116. Van der Perk M, Burrough PA, Voigt G (1998) GIS based modelling to identify regions of Ukraine, Belarus, and Russia affected by residues of the Chernobyl nuclear power plant accident. J Hazard Mater 61:85–90
117. Gillett AG, Crout NMJ, Absalom JP, Wright SM, Young SD, Howard BJ, Barnett CL, McGrath SP, Beresford NA, Voigt G (2001) Temporal and spatial prediction of radiocaesium transfer to food products. Radiat Environ Biophys 40:227–235
118. Jackson D, Wragg SK, Bousher A, Zeevaert T, Stiglund Y, Brendler V, Hedemann Jensen P, Nordlinder S (1999) Establishing a method for assessing and ranking restoration strategies for radioactively contaminated sites and their close surroundings. Nucl Energy 38:223–231
119. Shershakov VM, Trakhtengerts EA (1997) Decision-making in an emergency by using computer analysis with dynamically changeable rules. Radiat Prot Dosim 73:141–142
120. ICRP (1973) Implication of commission recommendations that doses be kept as low as readily achievable ICRP Publication 22. Pergamon Press, Oxford
121. ICRP (1977) Recommendations of the international commission on radiological protection ICRP publication 26, Annals ICRP 1(3). Pergamon Press, Oxford
122. Lochard J, Schneider T (2001) Application of decision aiding technologies for relocation strategies. In: Van der Stricht E, Kirchman R (eds) Radioecology: radioactivity and ecosystems. Fortemps, Liege, pp 504–536
123. IAEA (2000) Restoration of environments affected by residues from radiological accidents: approaches to decision making IAEA-TECDOC-1131. International Atomic Energy Agency, Vienna
124. Jacob P, Fesenko S, Bogdevitch I, Kashparov V, Sanzharova N, Grebenshikova N, Isamov N, Lazarev N, Panov A, Ulanovsky A, Zhuchenko Y, Zhurba M (2009) Rural areas affected by the Chernobyl accident: radiation exposure and remediation strategies. Sci Total Environ 408:14–25
125. Salt CA, Culligan Dunsmore M (2000) Development of a spatial decision support system for post-emergency management of radioactively contaminated land. J Environ Manage 58(3):169–178
126. Sazykina TG, Kryshev II (2006) Multi-criteria analysis for evaluating the radiological and ecological safety measures in radioactive waste management. Communications of Higher School. Nuclear Power Engineering (Izvestia vissikh uchebnikh zavedenij. Yadernaya energetika), Obninsk 1, pp 39–46 (in Russian)
127. Shaw G, Robinson C, Holm E, Frissel M, Crick M (2001) A cost-benefit analysis of long-term management options for forest following contamination with ¹³⁷Cs. J Environ Radioactiv 56:185–208
128. IAEA (2009) Quantification of radionuclide transfers in terrestrial and freshwater environments for radiological assessments IAEA-TECDOC-1616. International Atomic Energy Agency, Vienna

Books and Reviews

- Brechignac F, Howard BJ (eds) (2001) Radioactive pollutants: impact on the environment and humans. EDP Science, Les Ulis
- Brown J, Hammond DJ, Kwakman P (2009) Generic handbook for assisting in the management of contaminated drinking water in Europe following a radiological emergency. EURANOS (CAT1)-TN(09)-02
- Ilyin LA, Gubanov VA (eds) (2004) Large radiation accidents: consequences and protective countermeasures. Izdat Publisher, Moscow
- International Atomic Energy Agency (2006) Environmental consequences of the Chernobyl accident and their remediation: twenty years of experience, Radiological assessment reports series. Report of the Chernobyl forum expert group “Environment.” International Atomic Energy Agency, Vienna
- Izrael YuA (2002) Radioactive fallout after nuclear explosions and accidents. Elsevier, Amsterdam
- Nisbet AF, Jones A, Turcanu C, Camps J, Andersson KG, Hänninen R, Rantavaara A, Solatie D, Kostianen E, Jullien T, Pupin V, Ollagnon H, Papachristodoulou C, Ioannides K, Oughton D (2009) Generic handbook for assisting in the management of contaminated food production systems in Europe following a radiological emergency v2. EURANOS (CAT1)-TN(09)-01. Available at <http://www.euranos.fzk.de>
- Scott YuA (ed) (2004) Modelling radioactivity in the environment. Elsevier, Amsterdam
- Shaw G (ed) (2007) Radioactivity in the terrestrial environment. Elsevier, Amsterdam
- Smith J, Beresford NA (2005) Chernobyl catastrophe and consequences. Praxis Publishing/Springer, Chichester
- Van der Stricht E, Kirchman R (eds) (2001) Radioecology: radioactivity and ecosystems. IUR, Fortmeps, Liege
- Voigt G (ed) (2001) Remediation strategies. Special Issue. J Environ Radioactiv 56
- Voigt G, Fesenko S (eds) (2009) Remediation of contaminated environments. Elsevier, Amsterdam

Environmental Geology, Introduction

JAMES W. LAMOREAUX

P.E. LaMoreaux & Associates, Tuscaloosa, AL, USA

As world population increases, greater demand is placed on finite resources that earth can provide. Geographic areas once judged inadequate or inappropriate for construction, water and wastewater management, or transportation routes, among other activities, are being considered for and being used for development.

Environmental geology is the discipline which brings together these issues in a holistic approach to land use planning and utilization. This section of the encyclopedia examines some of the problems, techniques, and solutions for managing resources so that a balance can be maintained between development and the environment.

Natural disasters are reported frequently and have varying impacts depending upon the closeness of the disaster to population centers and the magnitude of the hazard. In all cases, the primary purpose of studying these phenomena is to minimize loss of human life and secondarily to minimize loss of infrastructure and impacts on present and future developments.

Each type of hazard impacts the environment in different ways, with some posing more danger to human life than others, and some posing more hazards to infrastructure and ongoing development. This section on environmental geology will address these various factors to provide a better understanding of how they occur, what can be done to minimize risk and how to plan better to address these concerns in the future.

Impacts of natural disasters including earthquakes and tsunamis and occurrence of sinkholes and landslides can have more devastating effects due to increases in world population. Two entries address earthquake issues: ► [Earthquake Faulting, Ground Motions and Deformations](#) and ► [Construction Planning, Environmental Impact of Foundation Studies and Earthquake Issues](#). Earthquakes have occurred throughout history and left their traces as manifestations of geological processes on the geological record.

Many lessons have been learned from damages caused by earthquakes. Guidelines have been developed

for design of earthquake resistant structures under different ground conditions. Education of the public has saved lives and continues to be a critical need as is research to develop mechanisms to forewarn the public.

Sinkholes and landslides are both naturally occurring and man induced. Both hazards can cause significant damages to various structures; however, they may not pose as significant a risk to human life as earthquakes. Protective measures can be performed to minimize the risks involved including preparing sinkhole maps which delineate the low, medium, and high risk areas subject to subsidence. If decisions are made to develop in sinkhole and landslide prone terrains, investigations of the geology and hydrogeology can determine where to seat structures into bedrock so as to minimize damages. In addition, engineering design can address the added stresses hazards pose and incorporate additional measures so that buildings and infrastructure can withstand them.

► [Karst Terrane and Transportation Issues](#) discusses various types of sinkholes, what their causes are and how to protect against their occurrence. This entry addresses ways to remediate damage and prevent future reoccurrence. ► [Land Subsidence in Urban Environment](#) reviews an emerging cause of subsidence in the burgeoning population centers of the world. Better management of water resources is required to prevent further subsidence from occurring in urban areas. Optimal utilization of water supplies must be implemented to correct existing situations and to minimize subsidence in the future.

Overuse of water supplies in concert with drought conditions and land degradation is creating desertification conditions rapidly in the world today. Desertification affects arid, semiarid, and dry subhumid regions; occurring on every continent, it impacts more than two billion people in these areas. Often these conditions occur across country borders and within areas of the world where tensions are already strained for political and religious reasons. ► [Desertification and Impact on Human Systems](#) discusses the activities of the United Nations Convention to Combat Desertification (UNCCD) and emphasizes the need for sustainable development at the local level. Stakeholders in these inherently fragile ecosystems must be brought

together to arrive at solutions that will allow continued production of ecosystem goods and services.

Not only do arid conditions require sustainable development, but urban areas do also. In *The Interrelated Roles of Natural Resources and Sustainability in Urban Planning*, cities are increasingly recognized as the most important global sustainability challenge of this century. Here too, stakeholders are critical in regard to resource access and distribution. Studies of urban metabolism are required to explore interactions among resource flows, urban transformation processes, waste streams, and quality of life. Results of these studies must be correlated to address the challenge of combining urban planning with sustainable resource management.

An integral part of continued growth and development of urban areas is detailed in ► [Groundwater Salinity Due to Urban Growth](#). Understanding of groundwater is critical in regions where there is an increase in water demand and where shortages in water supply affect available and proposed infrastructure. Knowledge of water accessibility in terms of quantity and quality can help reduce the risk of saline intrusion in coastal areas from urban growth. Socio-economic and environmental development programs require adequate political decision making and planning at local and regional levels. Proper space planning implies that proposed land use is in agreement with what land “attitude” and its environment can provide without putting in danger sustainability.

Geochemical modeling is applicable to investigations of water quality, as well as to many other fields. Modeling is a powerful and indispensable tool for research and investigations of environmental sustainability, science, and technology. It provides quantitative evaluation of complex processes and it can predict the extent and consequences of geochemical reactions in the order of tens of thousands of years. Basic concepts, information for accessing modeling codes and further readings are included in this chapter. More importantly, applications in the field are provided. An excellent example of the utility of geochemical modeling in environmental sustainability, science, and engineering is illustrated in geological carbon sequestration.

Much research is currently being performed on injection of carbon dioxide (CO₂) into deep geological formations as a climate mitigation tool. ► [Geologic](#)

[Carbon Sequestration: Sustainability and Environmental Risk](#) provides an overview of the research and ideas about the future. National and local regulations will require risk assessments of wellbore integrity, well injectivity, and long-term performance in the order of thousands of years. The geology of each site differs and performance assessments are necessary at all stages of CO₂ storage operations: site assessment and selection, design, installation, operations and monitoring, and closure and post closure.

Drilling into deep geological formations is also taking place offshore in an activity known as Oceanic Hydrocarbon Investigation (OHI). OHI is currently increasing its production volume in an energy hungry society and the quest for finding and exploiting underground hydrocarbon resources is being continuously improved.

This is particularly true in relation to underwater mapping, inspection, and monitoring of the environment. ► [Marine Life Associated with Offshore Drilling, Pipelines, and Platforms](#) states that OHI has inflicted oil-spills on marine and coastal environments. At large, however, OHI appears to be environmentally friendly. This notion has been documented by extensive seafloor mapping and annual visual inspections of platforms, pipelines, and other infrastructure.

Environmental concerns in marine environments are also being addressed as they relate to dredging of navigable waterways. Dredges of various designs have been used for many years to create and maintain navigable waterways to move people, goods, and materials. ► [Dredging Practices and Environmental Considerations](#) discusses the history of dredging from the times of the pyramids when people used long-handled dipper shovels, to when productivity gains from using animals increased digging power. In the 1800s development of electric and steam power units enabled construction of mechanical, backhoe, and pipeline dredges. Hydraulic technology in the 1960s, utilizing hydraulic winches and hydraulic rotary cutter drives, facilitated removal of finer grade sediment. As technology improved, so too did concern for the environment, such that today there is much debate and research on the impacts of dredging. This concern has heightened with contaminated sediments in the picture and the question of whether to dredge or leave them in place.

Navigable waterways integrally relate to dam construction and movement of people, goods, and services. Construction of dams has a much greater impact on the environment. Since time immemorial, people have looked for ways to tame surface waters to prevent floods and use water for different purposes, including transportation, irrigation, water supply, and electricity production. ► [Dam Engineering and its Environmental Aspects](#) details the history of dams and their primary role to store or to divert waters. Many dam projects have been controversial and potentially disastrous. Failure of dams and their environmental impacts are the main reasons for controversy and fear. This is particularly true in karst environments which tend to be more sensitive to the influence of dams and reservoirs. With increasing demand on water resources and electric power, an important issue is how to maintain the balance between the necessity for development and preservation of the environment.

Energy continues to weave a thread through this section of the encyclopedia as the chapter on mining deals with mining of coal. Also addressed are impacts from mining of rare earth minerals, gold and copper among others. ► [Mining and Its Environmental Impacts](#) details the demands of the growing world population for increasing amounts of raw materials. To find acceptance and support in society, however, any future mining activity will demand state of the art environmental management and must incorporate sustainable development. A history of mining is provided, as well as an overview of the environmental effects of mining, differentiated by its relevant phases: exploration, exploitation and processing, decommissioning and rehabilitation.

Nuclear energy has its own inherent concerns which are increased in times of natural disasters. ► [Groundwater Impacts of Radioactive Wastes and Associated Environmental Modeling Assessment](#) discusses the long-term risks to human health and environment from radioactive waste contamination in soil and groundwater. Major sources of radioactive wastes and their impact on groundwater contamination are reviewed along with the major biogeochemical processes that control the transport and fate of radionuclide contaminants. The evolution of mathematical models designed to simulate and assess the transport and transformation of radionuclides are

also described. Understanding the fate and transport of radionuclide contaminants can improve the ability to forecast contaminant destination and select cost effective remediation technologies.

Perhaps the greatest energy source known to man is from volcanoes. ► [Volcanoes of Mexico](#) includes an introduction to volcanoes: the reasons they form, where they are located, and the hazards presented by different types of eruption. Mexico is one of the world's most volcanic regions and a summary of Mexican volcanoes is presented. Volcanic eruptions can be divided into two broad categories: explosive or effusive. Various factors combine to determine how the magma emerges, the key ingredient of which is water. With the exception of ballistics, which do not reach very far from the volcano, volcanic hazards can present a major risk to populations living close by. Volcanologists need to investigate deposits of previous eruptions to understand what might happen in the future. Hazard maps can be created which indicate where eruptive products might accumulate.

Water, energy, minerals, and the environment all play a role in development and future growth of the world and its population. No single aspect can be addressed without including another. This section of the encyclopedia on environmental geology places these different factors into perspective so that scientists, politicians, economists, planners, and stakeholders can work together to develop solutions that provide an acceptable balance to development and the environment.

Environmental Impacts of an Open Ocean Mariculture Operation in Kona, Hawaii

NEIL ANTHONY SIMS

Kona Blue Water Farms, Inc, Kailua-Kona, HI, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Overview of the Kona Blue Water Farms Operation

Farm Operations
 Localized Environmental Concerns and Actual Impacts
 Global Impacts: Fish Meal and Fish Oil Usage
 Future Directions
 Bibliography

Glossary

Sustainability The use of resources in such a manner as not to impinge upon the ability of future generations to enjoy similar use.

Open-ocean mariculture The culture of marine fish, invertebrates, or algae in exposed ocean locations.

Definition of the Subject

Open-ocean mariculture is the culture of marine fish, invertebrates, or algae in exposed ocean locations. The criteria for demarcation of open-ocean sites from inshore or nearshore aquaculture sites is often the subject of debate, with various schemes using some integration of distance from shore, depth of water, or exposure to open water and high sea conditions. A more meaningful definition is perhaps, simply, that an open-ocean mariculture site is a site where any of these conditions or criteria combine to disconnect the culture system from the surrounding substrates that might otherwise be subject to some significant environmental impact, or might become reservoirs for some ecological feedback with the culture system. Alternatively, rather than using a geographical or operational distinction, open-ocean mariculture might simply be described as culture of marine animals or plants at sites that are further offshore, or in deeper water. This definition reflects the evolving, aspirational, and incremental improvement in culture systems, rather than thinking of the open ocean as a final destination.

Whatever the definition, open-ocean mariculture offers the potential for a scalable production system for valuable, healthful seafood, with negligible environmental impacts.

Introduction

The Seafood Crisis

The oceans are in deep trouble. Even though the demand for seafood has increased, capture fisheries

around the world are collapsing from overfishing, or are static. Wild stock fisheries cannot sustain any greater pressures, and clearly cannot scale to meet the growing needs of increasing population size, increasing affluence, and wider recognition of the health benefits of seafood.

The USA is both symptomatic of the seafood crisis and significantly contributive to the problem. In the USA, closures or buyback schemes to reduce effort have effectively shut down once-productive fisheries for Atlantic tunas and swordfish, the groundfish of Georges Bank and other Northeast fisheries, Pacific Coast sardines, albacore, and more recently, rockfish. Other environmental concerns for endangered species or marine mammals have seen closures or limitations placed on fisheries for shrimp in the Gulf of Mexico, purse seining for tuna in the Pacific, and long-lining for tuna and swordfish in Hawaii and the US Pacific. Currently, over 80% of the seafood consumed in the USA is imported, and more than half of those imports are from farmed sources.

Hawaii Fisheries Status The Hawaiian Islands represent a microcosm of the global fishing crisis. Increased fishing power – in terms of both the number of boats and available technologies – has seen the valuable deep bottom fish catches decline precipitously over the last few decades. GPS units allow deep bottom fishermen to return repeatedly and precisely to a “hot spot” until the fish are gone. Electric reels reduce the labor of hauling up from 60 fathoms to simply flicking a switch. Powerful echo sounders allow tuna fishermen to identify the location of individual fish tens of fathoms beneath their boat. At the same time, catches of sashimi-grade tuna have plummeted due to competing gear-types (seiners, long-liners, pole-and-line) and conflicting national interests throughout the range of these highly migratory species. Some 20 years ago, over 25 commercial fishing boats worked out of the local Honokohau Harbor in Kona, Hawaii, targeting deep-water snappers and big-eye tuna in the “ika-shibi” fishery. Today, there is only one commercial fishing vessel based there, and even that is only part-time.

The decreasing catch volume and decreasing average size of fish caught in the bottom fish fishery in Hawaii is causing increasing concern. By 1996, only 20% of the onaga (*Etelis coruscans*) catch in the main

Hawaiian Islands (MHI) had previously spawned; similar declines were evident among other species stocks. These species' biological characteristics make them vulnerable to recruitment overfishing; NOAA Fisheries staff estimate that an onaga attains maturity at about 4.1 years of age, at a size of 66 cm. With Federal and State data indicating significant overfishing of these stocks, increased regulation became imperative. In June 1998, new legislation went into effect, establishing limits on fishing gear, bag limits, registration of bottom fishing craft, and restricted fishing areas (up to 20% of the bottom fish ground was placed off limits) for the commercial and recreational fishery industry.

Further restrictions are still needed. Over the last two summers, the State declared all bottom fishing closed for the summer period from May 1st to September 30th, throughout the island range. State authorities have indicated that this seasonal closure may need to be repeated for coming years, as well, before there is any measurable improvement in stocks. The majority of high-value species consumed in Hawaii already are imported from other areas, such as the South Pacific and South East Asia.

Aquaculture as Part of the Solution Aquaculture offers the only viable solution to the growing demand for sustainable, healthy sources of seafood protein for human consumption. Fish farming reduces exploitative pressure on already depleted wild stocks, supports the growth of coastal and rural industries, and yields a product that is low in saturated fat and high in protein. The annual contribution of aquaculture to global aquatic production is now almost equal to that of wild catch (47% versus 53% [1]). In 1985, aquaculture represented only 5% of US fish consumption, yet today that figure stands at around 40%. Aquaculture growth is rapid, and is projected to increase in pace. The US Department of Commerce has set a goal of a fivefold increase in domestic aquaculture production value, to \$5 billion, by 2025.

Domestic aquaculture production using existing methods or species cannot keep pace. Almost all US production is from freshwater species; the only marine species cultured in any quantity are salmon and striped bass, both of which are anadromous (freshwater spawning). It is almost impossible to obtain permits for nearshore farm sites; there is intense competition

among different user groups for nearshore waters, and aquaculture has been besmirched in the conventional wisdom as environmentally destructive and unsustainable.

The Open-Ocean Potential Open-ocean mariculture technology has recently moved from the realm of science fiction [2, 3] to commercial reality. In the last few years, there have been dramatic advances in the legal and engineering fields, which have opened up the new fish farming frontier of offshore areas. New submersible net pen systems have been pioneered by OceanSpar, LLC, of Washington State ("Sea Station™" net pens), and Ocean Farm Technologies, Inc., of Maine ("Aquapod™" net pens). These new technologies have dramatically increased the workable extent of ocean farming, by providing seaworthy platforms for grow-out of fish in exposed offshore environments. These farms are therefore able to be located further offshore, in sites where currents, prevailing sea conditions or seasonal storm events may make surface pens inadvisable. These operations are also usually portrayed as having less environmental impact than nearshore fish farms. The rationale is that there is a cleaner and clearer disconnect between the farm and the underlying benthic substrate or the adjacent shoreline so that there is both minimal environmental impact and negligible potential for negative accumulative feedback from the environment to the culture system.

But how have these first open-ocean mariculture operations actually performed? Is there any basis in fact to these claims of environmentally sound mariculture in offshore locations? What are the opportunities for expansion of this industry, and what are the risks and precautions – both environmental, biological and economic – that should accompany such scaling?

This entry reviews some of the recent progress toward realizing this potential, as exemplified by one farm operation, in Kona, Hawaii. The environmental performance over the past 5 years of operation at Kona Blue Water Farms' site is instructive. It may demonstrate how open-ocean mariculture may represent a partial solution to the seafood crisis: a scalable, environmentally sound production system for high-value seafood. The environmental impacts of this one open-ocean mariculture site are reviewed here, in the context

of both local ecosystem effects and global marine resource use efficiency: the “footprint” on the oceans.

Overview of the Kona Blue Water Farms Operation

Prior to 1998, Hawaii’s ocean-leasing legislation limited any potential project to a maximum of 4 acres, which had to be used for either educational or research purposes, and not for commercial gain. Through several years of work by industry aspirants, and strong leadership by the State Aquaculture Development Program, legislation was passed that allowed commercial offshore fish farms or energy projects.

The Kona Blue Water Farms principals had been involved through the legislative review, and with the passage of the bill, began research into developing hatchery culture techniques for high-value marine fish, simultaneously surveying the Kona Coastline for prospective offshore farm sites. After an extensive 3-year process of consultation and consensus-building with the community, Kona Blue was granted the requisite State and Federal permits for the original offshore farm site in March, 2004. The operation began deployment in February 2005 and first fish were harvested offshore in September, 2005. Since then, production has grown to the extent that Kona Blue has been harvesting up to 25,000 lb of sashimi-grade Kona Kampachi® per week, and up to 500 T per year. Kona Kampachi® (*Seriola rivoliana*) is also known as kahala, Long-fin Amberjack or Almaco Jack. It is related to the Japanese hamachi (*S. quinqueradiata*), but is native to Hawaii, and is distributed throughout the warm waters of the world.

The Offshore Farm Site and Farm Operations

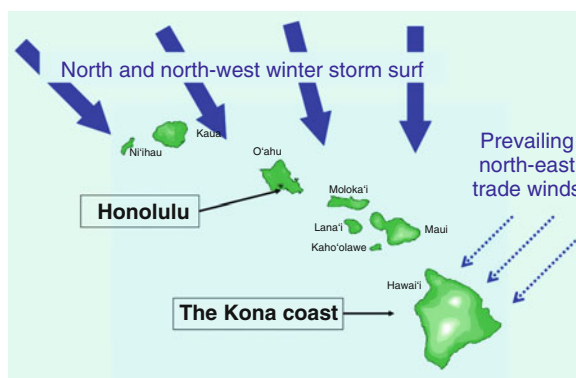
The Lease Area Site selection is a critical component for any mariculture operation, but is particularly so for an innovative offshore farm that is pioneering both a new permitting process and a new net pen system. The original farm lease site was selected on the basis of the following criteria:

1. The selected site was in a deep-water area, over 200 ft deep, with brisk currents.
2. There was little or no public use of this area. The farm site lay between the limits of normal

recreational scuba diving (around 120 ft) and the normal depths for offshore trolling for ono (wahoo, *Acanthocybium solandri*).

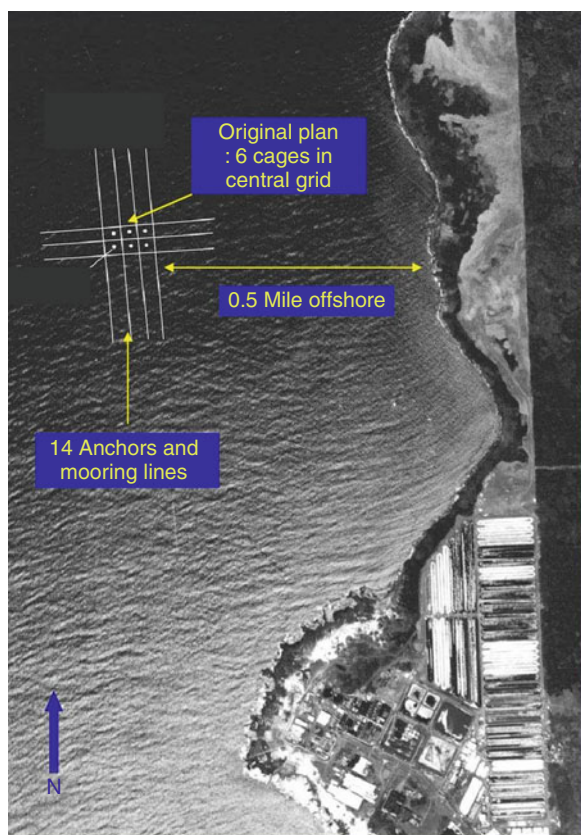
3. The site afforded some protection from both Kona storms and the strong trade winds (Figs. 1 and 2). The proximity to shore also allows for future telemetry links to shore for farm control and security.
4. There was ready access from Honokohau Harbor, 5 miles to the south, which provides support facilities such as slips, fueling, and land for staging of equipment and feed.
5. The site was directly offshore from the Kona International Airport and NELHA (Fig. 2), and, as such, its use was consistent with the adjacent land uses, and it represented no significant impact on the viewplane.

The farm site’s topography and oceanography are distinguished by the depth of water, the bare sand substrate, the strong currents through the area, the exposure to high winter surf and strong trade winds, and the adjacent shoreline of a narrow coral bench reef with a steep basalt (lava) cliff. A few black sand beaches also lie along the coastline, to the north of the site, but these are little used, except by recreational fishermen. The preexisting uses of the farm lease area itself were negligible, because of its depth, the paucity of fish, and the barren benthos.



Environmental Impacts of an Open Ocean Mariculture Operation in Kona, Hawaii. Figure 1

Location of the Kona Coast on the western side of the Big Island of Hawaii affords moderate protection from trade winds and winter swells



Environmental Impacts of an Open Ocean Mariculture Operation in Kona, Hawaii. Figure 2

Kona blue water farms' site located in waters over 200 ft deep over a sand bottom, a 1/2-mile offshore from a pristine coral fringing reef abutting lava cliffs. The site is a mile north of Keahole Point and west of the Kona International Airport

The net pens are all concentrated toward the center of the lease area (see Figs. 2 and 3), within two mooring arrays: one containing six net pens, and the other containing two net pens and the feed barge. The closest distance from the edge of the central grid array to shore is approximately 2,600 ft, or almost half a mile to the northeast, to Unualoha Point.

The farm site lease provides “negotiated exclusivity”: Transit, trolling, hoop-net fishing and hook-and-line fishing are permitted throughout the lease area, but for liability, insurance, and safety reasons, there is no authorized anchoring, scuba diving, or swimming permitted.

The 90 acre lease area initially accommodated eight submersible Sea Station net pens, each of around 3,000 m³ capacity (Fig. 4). The outermost area of the lease is used almost solely for mooring lines, which require a 5:1 scope. The net pens were originally tied into submerged grids that were anchored into the soft substrate using steel embedment anchors and chains. A series of buoys and weights ensure that the anchor lines are perpetually taut, to eliminate any risk of entanglement by marine mammals. Bridles from the mooring grid corners attach to the net pen rims, to hold the net pens in place in each grid square.

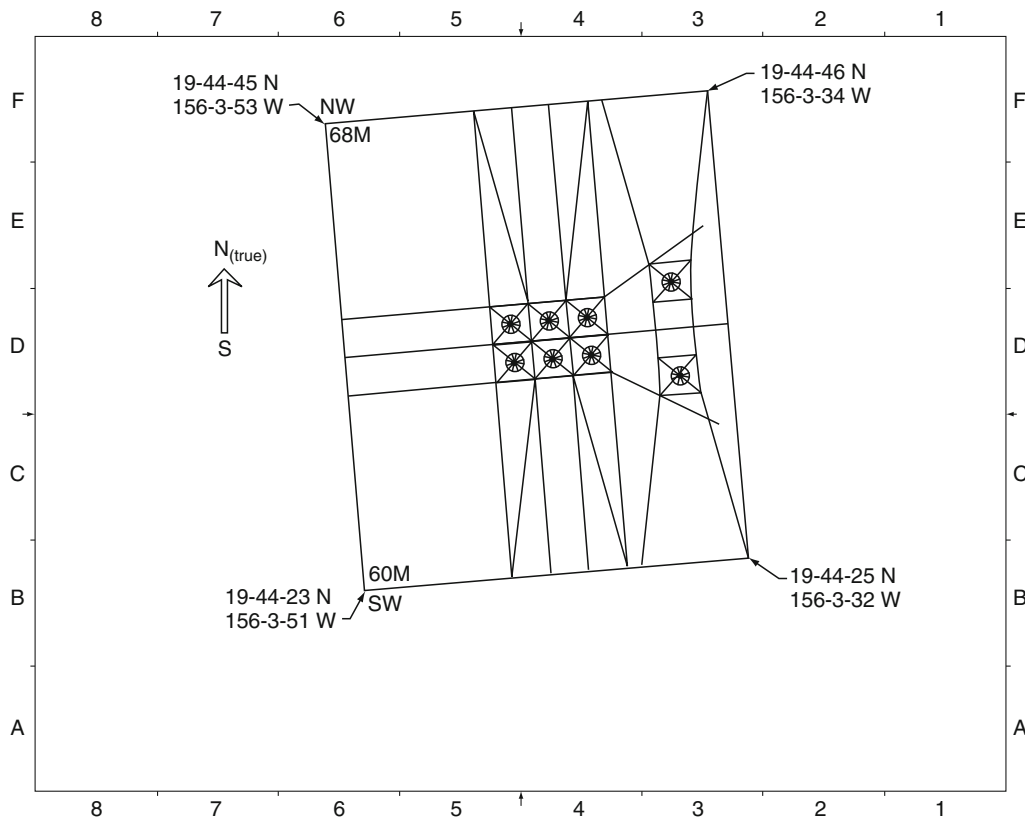
Farm Operations

The daily activities on the farm primarily consist of feeding the fish in the pens. Underwater video cameras inside the net pens are used to relay visual images to the operators on the feed barge. This enables the feed operators to regulate feed to ensure that no feed is wasted, and that excess feed does not fall below the net pen. Any fish carcasses are regularly removed by divers. Carcasses are disposed of as solid wastes in the county landfill.

Harvests usually occur twice each week. Fish are harvested into an ice-brine slurry, to quickly and humanely kill the animals with a minimum of damage. Fish are all transported whole, in ice-brine, to a single land-based processing facility, for packing and shipping. No fish processing occurs at sea during the harvests. Disposal of processing wastes is the responsibility of the wholesalers or other purchasers of the fish, but, at present, most trimmings from fillets go into the landfill.

Support activities for the existing operation are based out of Honokohau Harbor, where a half acre of land rented from the State accommodates containers for feed storage, gear storage areas, a closed workshop area, restroom, and office.

The farm is also serviced by a semipermanent feed barge/security platform vessel, which has been deployed on-site since October, 2007. A separate harvest boat – the 74 ft F.V. Kona Kampachi – transports the harvested product back from the farm site to the harbor. Several other smaller work boats are also used to support net pen and grid maintenance and cleaning, and other tasks.



Environmental Impacts of an Open Ocean Mariculture Operation in Kona, Hawaii. Figure 3

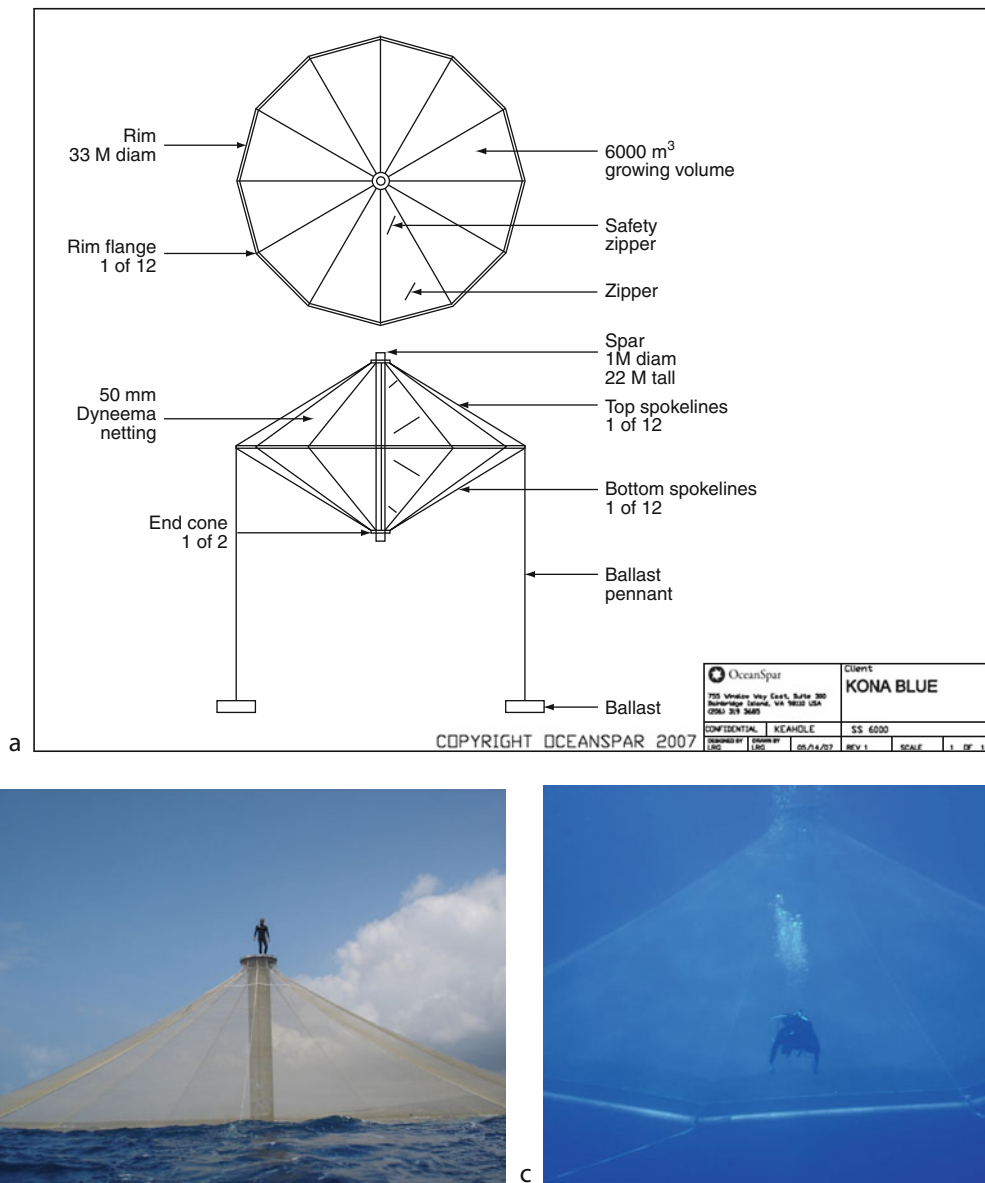
Modified mooring array and grid dimensions – plan view. The number of net pens is presently being reduced from current eight smaller pens to five larger pens, with the same overall culture capacity ($24,000 \text{ m}^3$). The submerged central grid remains at around 30 ft (9 m) beneath the surface

Localized Environmental Concerns and Actual Impacts

The Presumed Problems

Aquaculture – or indeed, development of any food production system – brings with it attendant environmental concerns. Fish farms are widely accused of environmental degradation. The concerns that are often voiced include the following local-affect possibilities:

- Detrimental impacts on water quality
- Nutrient enrichment of the substrate beneath the farm
- Antifoulant paints from net pens to contaminate the substrate
- Therapeutant or antibiotic misuse to harm the surrounding biota
- Escapes to outcompete wild fish for spawning grounds or feed
- Escapes to dilute the wild fish gene pool
- Proliferation of pests, parasites, and diseases inside the net pens, which can then be transferred to wild fish
- Entanglement of whales, dolphins, and other marine mammals
- Disruption of marine mammal or other species' migratory paths
- Harmful deterrents or fatal control measures against predators
- Excessive use of fish meal and fish oil, leading to overharvesting of the smaller pelagic species targeted by industrial reduction fisheries



Environmental Impacts of an Open Ocean Mariculture Operation in Kona, Hawaii. Figure 4

Submersible Sea Station® Net Pens deployed on Kona Blue farm site. **(a)** Design of submersible SS6200 Sea Station net pens, with central steel spar and steel rim. **(b)** Sea Station SS3000 raised to rim-level on the Kona Blue farm site. **(c)** A submerged Sea Station SS3000 and diver on Kona Blue site

- Exclusion of other user groups from traditional, cultural, or recreational uses of the farm area
- Visual impact of the net pens on the viewplane

With almost 5 years' experience at the Kona Blue farm site, then, it is appropriate to evaluate the actual data and observations recorded at the Kona operation,

and to compare this experience with the concerns that had been, and continue to be, voiced. Each of these issues is examined in detail, below, beginning with an evaluation of the de novo environmental status of the Kona Blue farm site, and then detailing the impacts that have occurred, their context, and their significance.

The Actual Observed Impacts: Locally

Water Quality and Effluent Impacts The water quality at the farm site is close to oceanic, with strong currents and low turbidity. Underwater visibility usually exceeds 100 ft or more.

General water movement patterns at the farm site are governed by the longshore currents past Keahole Point (the western-most point of the Big Island of Hawaii), 1 mile to the South. An S4 current meter deployed at the farm site over several periods since 2004 showed regular peak current speeds of over 50 cm/s (about 1 kt, at a depth of around 40 ft). Current headings were longshore: generally to the North (predominantly), but also to the South. The two points of first impact downstream from the farm site are therefore either Keahole Point, around 1 mile to the south of the site, or the Mahai'ula-Makalawena shelf area, around 3 miles to the north.

Because of the community concerns about potential impacts from the farm operation on water quality, the company had made commitments during the original permit process to a policy of ongoing transparency and objectivity in monitoring. These commitments included:

- Use of objective, third party experts to collect the water quality samples.
- Use of local water quality laboratories – such as NELHA Water Quality Lab, or local private laboratories – for conducting the sample analysis.
- Placement of copies of all monthly water quality monitoring reports at local repositories, such as the State Aquatic Resources office at Honokohau, or the NELHA library, so that local residents can review this data.
- Provision of reasonable access to Federal, State, and County officials for monitoring and oversight purposes.

Monthly measurements have been taken of ammonia and turbidity (the two most relevant water quality parameters for fish farming) at three depths (surface, mid-water – 50 ft deep, level with the submerged net pens, and at the bottom) and at a total of seven stations (two control stations upcurrent, one effluent station immediately downcurrent of the net pen with the greatest biomass, and four “zone of mixing” stations

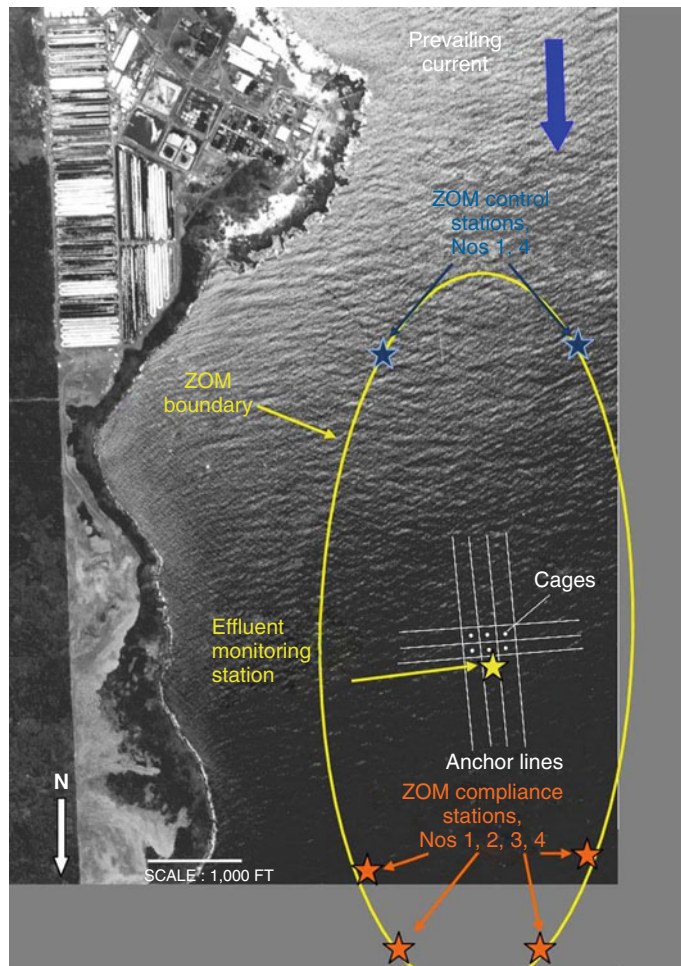
4,000 ft downcurrent: Fig. 5). Quarterly measurements are also taken for a range of other parameters.

Detailed water quality data are available on the company's web site [4]. Figure 6 shows the mean for each sample site for turbidity for September, 2008, when the farm was at peak production of around 500 T annually. Turbidity is probably the best metric for fish feces and other particulates in the water, and so is most likely to reflect any impact from the farm's presence. These data are definitively clear – there is no discernible difference between water quality parameters at the upcurrent control sites, and the effluent site (1 m downcurrent of the net pen with the highest biomass) or the “zone of mixing” sites downcurrent. These results confirm that *there is no measureable impact* on water quality from the existing farm operations.

Benthic Impacts The substrate beneath the farm is over 200 ft deep, and almost exclusively comprises bare, coarse sand. Located along the shoreline, some 2,000 ft to the East (directly across the longshore currents) is a diverse coral reef community.

Impacts on Substrate Beneath and Around the Farm Site Prior to farm installation, a preliminary survey of the site was undertaken by repeated bounce dives, using scuba, to depths of 220 ft. Because the depth of the farm site is beyond the limits of normal safe diving, and the strength and unpredictability of the currents precluded ready use of grab samples or drop video cameras, the original permit provided that no benthic monitoring would be required. Over time, however, permit requirements were tightened to include grab sample monitoring of substrate chemistry and infaunal micromollusk community structure, and video monitoring using drop cameras. These samples have been conducted quarterly, and reports and video footage have been posted on Kona Blue's website [4].

These results generally indicate that there has been no measureable impact on the benthic community around the farm site. There have been episodic perturbances of substrate chemistry immediately underneath the cage footprint, with a few instances of anoxic conditions during 2007, during periods when a new feed distribution system was being tested. This resulted in some pulverization of pellets, and reduction of feed to a “slurry,” rather than discrete pellets. Once



Environmental Impacts of an Open Ocean Mariculture Operation in Kona, Hawaii. Figure 5

Water quality sampling station map. Aerial photograph of the Kona Blue site showing water quality monitoring sampling station locations. Sampling stations under a prevailing N-setting current are shown. Under a S-setting current, control and compliance stations will be reversed. ZOM = Zone of Mixing

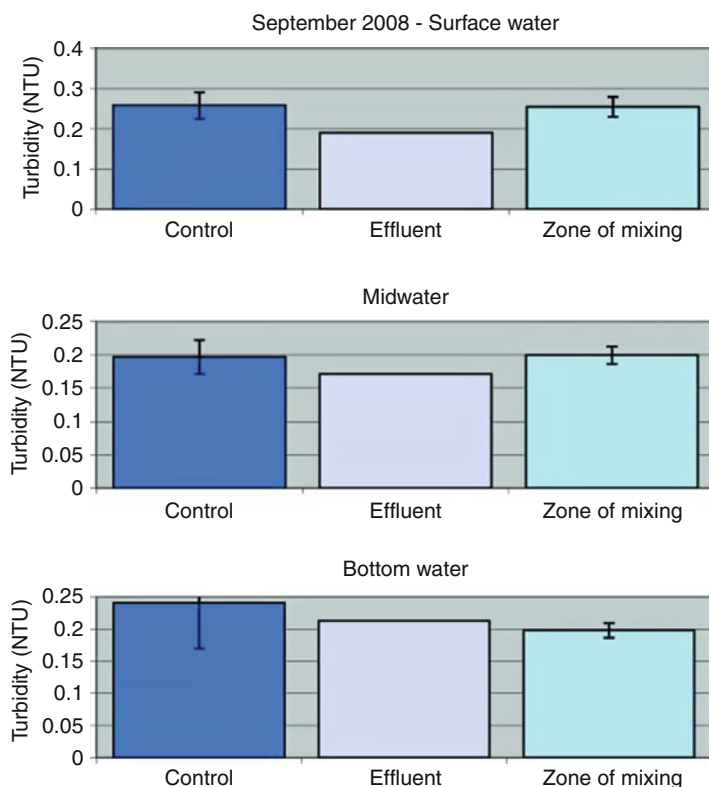
the feed system was refined, the substrate returned to its more normal condition, there was no further significant nutrient enrichment of the substrate.

Filamentous algae have also been visible in the drop-camera videos from around the farm site. These appear to have been detached from the cage mesh or the mooring lines, as the algae are not attached to the coarse sand substrate. Presumably, these algae are dispersed during periods of high current.

Monitoring of infaunal micromollusc assemblages in the substrate samples has also demonstrated that there has been no significant change in

the community structure resulting from the farm presence (see reports on-line: [4]).

Impacts on the Adjacent Coral Reef Community A comprehensive survey of marine biota was conducted on the reef directly adjacent to the existing farm lease area, just south of Unualoha Point [5]. The survey of the benthic biota of the fringing reef crest used protocols identical to those employed by the State's Division of Aquatic Resources, in their West Hawaii Reef Management Task Force Survey. This provided an extensive set of "control" sites: the other benthic and



Environmental Impacts of an Open Ocean Mariculture Operation in Kona, Hawaii. Figure 6

Typical turbidity data from monthly sampling around the Kona Blue farm site. There is no discernible difference in water quality between the two Control sites (upcurrent), one Effluent site (immediately downcurrent) and four Zone of Mixing sites (4,000 ft downcurrent) of the farm site

fish data from the sites along the 90 miles of coastline on West Hawaii. A series of four transects of 25 m × 2 m extended parallelly to the reef crest, immediately shoreward of the seaward edge of the reef. Video footage was made of these transects, and digitized for selection of random points on the video frames.

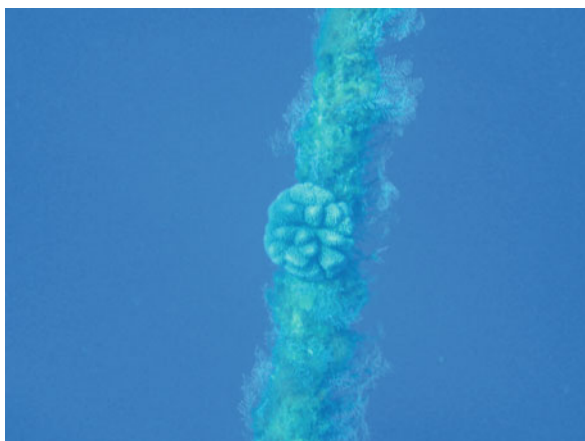
The Makako Bay–Unualoha site has been repeatedly resurveyed since the original 2003 survey. Although no formal reports have been compiled, there have been no significant changes in benthic community composition or fish populations reported.

Biofouling on the Farm Structures There is also profuse growth of macro-invertebrate biofouling on the grid-lines and buoys of the mooring array, as well as on the bridle lines that attach the cages to the grid and the rims of the cages themselves (Fig. 7). This fouling includes diverse macroalgae, bivalves (several species of mussels and oysters: *Pteria* sp and *Pinctada*

spp), corals (primarily *Pocillopora* and *Porites*), sea urchins (primarily *Echinothrix calamaris*) nudibranchs (*Stylocheilus longicauda*) and sponges. These all settle out of the plankton onto the farm structures, and their presence does not represent any significant or even measureable reduction in the available recruits to the nearby coral reef area. The growth of the corals, particularly, is compelling evidence that the presence of the fish farm operation is not deleterious to benthic organisms.

Apart from the one brief instance of anoxic conditions beneath the net pens, there have been no other adverse impacts on benthic communities in, underneath, or around the net pen area.

Pests, Parasites, and Pathogens Kona Blue employs an integrated pest management strategy to optimize fish health, reduce interactions or minimize impacts



Environmental Impacts of an Open Ocean Mariculture Operation in Kona, Hawaii. Figure 7

A *Pocillopora damicornis* coral colony on a mooring grid line around the Kona Blue net pens (within about 15 m of net pen stocked with fish). The presence of coral colonies that are highly sensitive to nutrient enrichment on moorings and buoys confirms that the operation has no significant impact on marine biota in the area

on wild fish stocks, and reduce any potential environmental impacts from therapeutant use. As *Seriola rivoliana* is a new species, any therapeutant use must be conducted under an Investigational New Animal Drug (INAD) permit. INAD permits operate under the oversight of US Fish and Wildlife Service (USFWS), and the Food and Drug Administration (FDA), with State oversight through Office of Conservation and Coastal Lands (within Department of Land and Natural Resources) and Clean Water Branch (CWB, within the Department of Health). Federal Environmental Protection Agency (EPA) has oversight through the NPDES (National Pollutant Discharge Elimination System), which is administered by CWB.

As with almost all farmed animals, *S. rivoliana* is subjected to small external pests – in this case, the skin fluke, *Neobenedenia* sp, that attaches itself to the fish's skin. These flukes do not pose any risk to human health, and do not themselves detract from the quality of the harvested product, but may cause irritation to the fish. If left unchecked, the flukes can become a health problem for the animal, as the fish rub themselves on the netting to ease the irritation. Kona Blue uses occasional treatments of dilute hydrogen peroxide

solution (at effective dosage rates of 200–300 ppm) to control levels of skin flukes among the fish in the net pens. Hydrogen peroxide (H_2O_2) breaks down rapidly in sunlight to form oxygen and water. Hydrogen peroxide is also considered an acceptable Organic aquaculture treatment under the draft USDA Organic aquaculture guidelines, and USDA Organic agriculture standards.

Under the permits in place at the existing site, such therapeutant use must demonstrate that there is no risk to the fish under treatment, or to the environment, or to human health. Monitoring of the effluent from any bath treatment at 100% concentration is mandated under the “Whole Effluent Toxicity” (WET Test) section of the NPDES permit. Results to date from the existing farm operation confirm that there are no significant environmental impacts from the use of the hydrogen peroxide. Ongoing effluent monitoring for WET test bioassays using larval fish (Pacific topsmelt, *Atherinops affinis*; conducted by Nautilus Laboratories in San Diego) demonstrate no significant difference in the rates of larval fish survival between control samples taken 4,000 ft upcurrent of the net pen, and samples taken of the whole effluent (100% concentration of the bath treatment water) at the conclusion of the bath treatments. There is therefore no mechanism for any measureable impact on the pelagic or benthic communities, or the surrounding water quality from the use of this therapeutant.

In addition, monitoring of wild kahala (*Seriola rivoliana*) stocks indicates that there is no significant proliferation of *Neobenedenia* sp. in the population around the farm area. Broodstock are collected periodically from around the farm area, to replenish the wild stocks in Kona Blue's hatchery. These fish are usually taken by commercial fishermen along the “drop off” of the marlin fishing grounds, about 1 mile to the South of the farm, and are sampled for ectoparasites upon capture, by immersion in a freshwater dip. Although these fish are usually infested with a number of other ectoparasites, the prevalence of *Neobenedenia* sp has never averaged much more than one individual per fish. By contrast, a parasitic copepod (sea lice, similar to *Caligus*) infests wild fish at average rates of around ten individuals per fish, and yet is not found at all on the farmed fish, and does not proliferate within the net pens.

A number of innovations, either in progress or planned, should also further reduce the proliferation of *Neobenedenia* on fish inside the net pens. The farm is being reconfigured to fewer, larger Sea Station net pens. With a planned reduction in the number of net pens, a reduction in the surface area-to-volume ratio of the remaining net pens (from double-cone net pens to a more cylindrical shape), the improved surface material characteristics and rigidity of the Kikkonet™ plastic monofilament net mesh (which make it easier to clean), and the improved access for offshore crew to regularly clean the nets from the surface (thereby breaking the skin fluke life cycle by dislodging the adhesive eggs on the mesh), the proliferative tendencies of the skin fluke should be further reduced.

Kona Blue does not use prophylactic antibiotics, but has, under the same regulatory oversights described above, and with veterinary guidance, used Florfenicol® to treat *Streptococcus iniae* infections that sometimes afflict juvenile fish after the stresses of transfer offshore. These treatments last for 10 days, and are also accompanied by WET test water quality monitoring. These WET tests have repeatedly demonstrated no impact on marine biota. A vaccine is available for *S. iniae*, which would avoid the need for these treatments, but regulatory requirements ironically prevent the use of the vaccine at this time. (The vaccine would qualify as an Organic treatment under draft USDA Organic aquaculture guidelines). *S. iniae* infections are not an issue with larger fish, once they have overcome the initial stress of transfer from the nursery to offshore.

Much of the concern over proliferative capacities for fish farm pests, parasites, or pathogens is derived from conflicts between salmon farming and wild salmon runs. Some research – though disputed – suggests that sea lice infestation rates can be exacerbated by the presence of salmon farms and can then be detrimental to survival rates of juvenile salmon as they migrate past the farms to the sea. Most marine fish, however, are broadcast spawners. Juvenile marine fish are therefore dispersed over vast areas of ocean and reef, and do not usually have vulnerable migratory patterns. Given such a distinct difference in life histories between salmonids and marine fish, there would seem to be limited applicability of the salmon and sea-lice research, or the concerns with impacts on vulnerable life stages, to open-ocean mariculture.

Interactions with Wild Fish Kona Blue cultures only Kona Kampachi® (*Seriola rivoliana*) on the offshore site, but the pertinent State permit also allows the company to possibly culture other amberjack (the other kahala species, *S. dumerili*), mahimahi (*Coryphaena hippurus*), and Pacific threadfin (*Polydactylus sexifilis*).

Aggregative Effects on Wild Fish Stocks The existing operation does have an aggregative impact on some species of fish in the area, but this is considered neither deleterious nor significant. Fish are attracted to the site for a number of possible reasons: the fouling on the net pen, the occasional release of small quantities of uneaten food from the net pen during periods of strong currents, and the aggregative nature of objects in open water (as for Fish Aggregation Devices). The makeup of the resident and transient fish communities around the net pens varies over time.

Pelagic or larger demersal fish frequently occurring around the Kona farm site include mackerel scad (opelu: *Decapterus macarellus*), ulua (giant trevally, *Caranx ignobilis*), wild kahala (*Seriola rivoliana* and *S. dumerili*), and barracuda (kaku, *Sphyrna barracuda*). Occasionally, schools of rainbow runners (kamanu, *Elegatis bipinnulatus*) and false albacore tuna (kawakawa: *Euthynnus alletteratus*) move through the net pen area. Larger pelagic fish, such as yellowfin tuna (ahi: *Thunnus alabacares*) and wahoo (ono: *Acanthocybium solandri*) are also occasionally attracted to the area by the baitfish, or by the net pens themselves.

A number of other, smaller fish that are more normally associated with coral reefs settle out of the plankton and assume residence either around the subsurface buoys or around the cages themselves. Such residents include schools of Sergeant majors (*Abudefduf abdominalis*), dascyllids (*Dascyllus albisella*), chromids (primarily *Chromis hanui* and *C. ovalis*) wrasses (primarily *Coris* spp and *Thalassoma* spp), and kyphosids (*Kyphosus* spp). As these fish are settled from the plankton, their presence is not considered a significant detractor from the biomass or diversity of the fish fauna on the adjacent reef.

Escaped Fish Interaction with Wild Stocks Concerns about potential negative impacts of escaped fish are often cited as one of the reasons for objections to fish

farming. However, this issue is most pressing only where non-native fish are cultured in areas where escapes might become established or compete with local species, such as Atlantic Salmon in the Pacific Coast of Canada. Kona Kampachi®, by contrast, is native to the waters of Kona. In addition, Kona Blue recognizes that the innovative net pen engineering employed means that there is some possibility of escape incidents over the initial proving period and development of refinements. In consideration of this, Kona Blue has deliberately not applied any selective breeding in the hatchery, and has not used any broodstock beyond F2 (i.e., all broodstock are either wild-caught, or first- or second-generation captive-reared). There is, therefore, no mechanism for development of any significant difference in the genetic makeup of the fish inside the net pen from the fish in the wild. This reduces any potential impact from escapes to merely direct ecological impacts.

Furthermore, the concerns with the effects of fish farm escapees on wild fish genetics are, again, largely a consequence of the conflicts between salmon-farming interests and wild salmon conservationists. Yet wild salmon stocks are unique, in that each river system or stream may have a genetically discrete stock from the adjacent watershed. Any blurring of this finer-scale differentiation, by interbreeding between escaped salmon and wild stocks, could represent a loss of genetic diversity. However, these concerns are not germane to farming of marine fish in the open ocean. As marine fish are broadcast spawners, there is only a coarse zoogeographic genetic granularity. Tagging research demonstrates that *Seriola* and other carangids migrate frequently between islands in the Main Hawaiian Archipelago. One *Seriola* migrated from French Frigate Shoals, in the Northwestern Hawaiian Islands, to the Big Island – a distance of 678 miles (over 1,000 km) over 3.6 years, at liberty [6]. The potential genetic impacts of Kona Kampachi® escapees on the wild stocks of *S. rivoliana* are therefore minimal.

Those Kona Kampachi® that have escaped from the Kona Blue net pens – either through “leakage” as divers enter or leave the pen through a submerged zipper, or from breaches in the netting – are invariably subjected to very heavy predation pressure. Individual escapees survive outside of the zipper for usually less than a minute before being eaten by either the ulua or the

bottlenose dolphins that are frequently in the area. The long-term prospects for survival and reproductive success of any escapees are therefore highly dubious. In addition, any escapees that do survive in the wild are presumably entering a wide-open ecological niche, due to the severe depletion of other deep-water species – such as the deep-water snappers – by commercial fishing. There is little likelihood of escapees competing in any significant manner with the few remaining wild snapper stocks.

Other Wildlife Interactions

Sharks The single overarching feature of shark interaction with the offshore fish farm site has been – contrary to conventional wisdom and activist concerns prior to the farm deployment – the general absence of sharks around the net pens. For the first 8 months of operation, only one fleeting shark sighting occurred: a small tiger shark (mano: *Galeocerdo cuvier*). Observations by farm workers suggest that there is a general pattern of brief influxes of tiger sharks to the area in the months of September and October of each year. Most of these animals appear individually, or in pairs, with a range of sizes from 8 to 15 ft in length, and they do not generally seem to take up residence on the farm site. Most tiger sharks only show interest in dead fish inside the net pens, and generally exhibit no interest in or aggression toward the farm workers.

In the first year of operation, however, a single tiger shark appeared to take up residence at the farm site. As the animal began to show aggression toward inanimate objects such as surface buoys, and then farm divers, it was humanely dispatched. Recognizing the long-term unacceptability of such predator control measures, Kona Blue sought alternative means of addressing this issue. A Shark Management Plan was then developed in consultation with State Aquatic Resources personnel in Kona and other experts, which included a range of measured responses, and nonterminal resolution if animals ever again become problematic at the site.

In subsequent years, tiger shark sightings usually increased in frequency at the farm site in the late-September early-October period. However, sharks were neither persistent, nor consistent. Farm operations had become more adept at removing dead fish, and the Shark Management Plan allowed divers to continue to work safely. One animal – or rarely two,

contemporaneously – may appear at the site, and remain for an hour or so, before moving away, presenting little inconvenience to farm operations, and no real risk to diver safety.

Kona Blue has also, in collaboration with DAR and HIMB researchers, established a receiver station on the farm site, as part of the larger research program for tracking tiger shark movements along the West Hawaii coastline. The first data series obtained suggested that the observations by the farm work crews were correct – that tiger sharks only very infrequently pass by the site, and rarely do they show any interest in the operation. From July 2006 to May 2007, there were a total of eight (8) records of tagged tiger sharks in the Kona Blue farm area. None of these sharks took up residence. One animal passed by the farm site three times in 2 months, another animal was recorded twice in 2 months, and three other animals had single records. (Fig. 8).

Over 2008 and 2009, however, further tiger shark tagging trials showed that two animals appeared to regularly return to the farm site over periods of up to 5 months. Two other sharks ranged over the entire Kona coast area, but for several weeks at a time were recorded exclusively from the farm site. All animals eventually moved on; one was later detected off Maui. While these results suggested that the farm site had become a “waypoint” for the animals over a few

months, the “long-term entrainment (e.g., years) of tiger sharks is unlikely” [7].

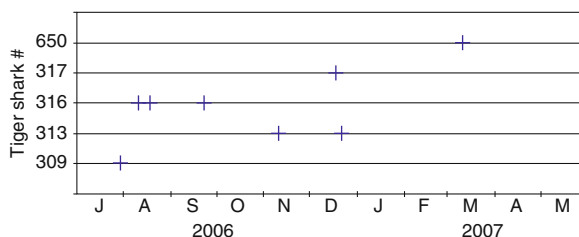
There have also been sightings of sandbar sharks (mano: *Carcharhinus plumbeus*) around the net pens. Initially, these were rare (none in the first year of operation), but since October, 2006, the frequency of sightings and number of sandbars has increased. These animals are usually seen in small groups (one to four sharks), below the net pens at depths of over 100 ft. They rarely rise up to the level of the net pens. Sandbar sharks are more secretive, and cannot readily be distinguished by any markings. No sandbar sharks were caught during the tagging trials in 2008–2009 [7]. It is therefore unclear if these are always the same individuals, or if they represent a larger population of animals that periodically move through the area.

In the period from June to August of 2008, there were a series of breaches of varying sizes in the Dyneema® webbing of one net pen that corresponded to shark bites. The same net pen was also breached in August 2009 by a small Galapagos shark that entered the net pen. The Galapagos was captured and released alive by company divers, unharmed except for a small dorsal fin notch for later identification. In each instance, breaches were sealed immediately on discovery. These incidents underscore the vulnerability of even sturdy Dyneema® nylon mesh, and have led to a plan for wholesale installation of Kikkonet® rigid plastic webbing across the farm. This material has been used in *Seriola* culture in Japan for over 25 years, and has been successfully used in crocodile and shark-infested waters by a sea-cage barramundi farmer in North Queensland, Australia. Kona Blue therefore anticipates that the use of Kikkonet webbing will reduce mesh breaches to negligible levels, and significantly reduce escapes and the attractant nature of the escapes to the bottlenose dolphins and sharks.

Overall, the evidence from the Kona Blue site confirms that there are no significant negative impacts from any aggregating effects of the net pens on sharks. The evolution of a nonterminal, humane plan for managing sharks on the farm site underscores the importance of commercial experience to improve open-ocean farming practices.

Turtles The threatened green sea turtle (*Chelonia mydas*) is common in the nearshore waters of the

Abacus plot showing dates (blue crosses) on which five transmitter-equipped tiger sharks were detected at Kona fish farming cages



Environmental Impacts of an Open Ocean Mariculture Operation in Kona, Hawaii. Figure 8

Frequency of tagged tiger shark occurrence at Kona Blue farm site: 2006–2007. Five tagged sharks were recorded over an 11-month period, with the most frequently occurring shark being present three times over a 2-month period. No animals took up residence, or showed any strong site affinity

main Hawaiian Islands. The endangered hawksbill turtle (*Eretmochelys imbricata*) is infrequently found in Hawaiian waters. The principal nesting site for the green turtle is in the Northwest Hawaiian Islands, on French Frigate Shoals [8]. No turtles have been observed in the area of the farm site, but it is possible that they occasionally transit through the site. If they were to do so, the taut-line mooring system and stiff-mesh net pens will prevent animals from becoming entangled.

Seabirds The submerged net pens used by Kona Blue do not significantly impact seabird populations. The farm area itself is infrequently used as a foraging area by seabirds. Most seabird activity in the area is confined to the fishing “grounds,” which extend to the northwest of Keahole Point.

Monk Seals There are four conceivable ways for open-ocean fish farming to have a significant negative impact on rare, threatened, or endangered wildlife, such as monk seals, dolphins, or whales. The project may (1) present a significant obstruction to natural migratory patterns, either (2) attract, or (3) repel the animals and thereby disrupt their normal behavior, or (4) the animals may become entangled in the ropes or mesh of the net pens or moorings.

Monk Seals have been observed at the existing farm operation on two occasions, both in association with escape incidents from the nylon mesh nets on the surface nursery pens that were previously in use at that site. (These nylon mesh surface net pens were removed in 2006, as Kikkonet was, at that time, not yet available outside of Japan.) On each of these occasions, the Monk Seal was preying on the small, escaped Kona Kampachi®, but once the school was effectively eradicated by predators, the Monk Seals moved away. A radio tag allowed movement of one monk seal to be tracked from the Unualoha site to a beach on Maui the following day, clearly affirming that the animal did not take up residence, or become conditioned to the availability of escapees.

Dolphins Makako Bay, almost half a mile to the south of the farm site, is frequented by large schools of spinner dolphins (*Stenella longirostris*), on nearly a daily basis. These animals usually follow a diurnal pattern of movement from the Makalawena shelf area to the

north, along the reef edge to the shallow areas of Makako Bay, where they rest for some time during the middle of the day. Some concerns were expressed during preliminary hearings about the potential for the farm operation to interfere with the spinner dolphin patterns of movement or resting habits [5]. There is no evidence to suggest that this has been the case. There have only been some occasions over the 5 years of operation offshore when divers or workers on the farm site have witnessed spinner dolphins coming anywhere near the net pens. The net pens clearly do not impede the usual pattern of spinner dolphin movement toward Makako Bay; they do not attract or repel the animals, nor do they affect the resting pattern of the dolphins.

Over the last 3 years, the existing farm operation has demonstrated a propensity to attract bottlenose dolphins (*Tursiops truncatus*). No bottlenose dolphins were previously present on the farm site, but the animals have begun to appear regularly at the site since about October, 2006. Patterns of dolphin movement are best characterized as one or two animals, every day or so, with occasional instances of groups of up to seven or eight animals. There is no regularity to the animals' appearance on the farm site: they may be present all day, or only in the morning, or only in the afternoon.

Kona Blue staff monitor and report on dolphin activity to HIHWNMS and NOAA's PIRO PRD. The bottlenose dolphins are probably attracted to the farm site by a combination of (1) the presence of the mid-water structures acting as a Fish Aggregating Device and the associated fish community that is present around the net pens, (2) the occasional provisioning from “leakage” escapes when divers enter or exit a net pen, and from the rare larger escape incidents when predators have breached the Dyneema nylon webbing, and (3) interaction with divers outside of the net pen, as the divers move about the farm from boat to net pen and back.

One individual dolphin has taken up residence over 2009 and 2010. This animal was suffering from a large fishing hook and leader line that had become lodged in its jaw, and it was present on the farm site almost continuously during this period. For many months, the dolphin was lethargic and lost weight, but more recently (as of late 2009) has appeared to be more active and in better condition [9]. The aggregative effective of

the net pens for this one animal might therefore be interpreted as beneficial.

No other individual bottlenose dolphin has taken up permanent residence at the farm site. There are no other animals present on the farm site on around one-quarter to one-third of days. Even when other animals are present, they are often only there for part of the day, rather than the entire day. In October–November, 2008, for example, dolphins were present for some period of time on 22 days out of 34 days [10]. There were dolphins present at the farm site for some or all of the day on 65% of the days. On 35% of days, there were no dolphins reported as observed on the site. Only on 1 day were six dolphins present. Most other days, there were one or two animals present for some portion of the day.

Other dolphin species may be found in and around the proposed farm lease area, but they are usually most commonly seen on the “grounds” to the south of the site. Spotted dolphins (*Stenella attenuata*), rough-toothed dolphins (*Steno bredanensis*), and false killer whales (*Pseudorca crassidens*) have all been observed on the “grounds,” or in other offshore waters of the Kona Coast, but have not been reported from the farm site.

In summation, although there has been behavior modification in one compromised individual, the presence of the farm operation has not had a significant negative impact on dolphin behavior. The overall long-term impact on dolphins from the farm operation will probably be further reduced. Modifications to net pens currently under way should help to alleviate the attractive nature of the farm to the dolphins, by reducing the potential for escapes through mesh breaches, and for leakage escapes, and by reducing the amount of time that divers need to operate outside of the net pens. Kona Blue will continue with the ongoing monitoring and reporting of marine mammal activity around the farm site, and continues to collaborate in this with HIHWNMS staff, PIRO PRD staff.

Humpback Whales Populations of the endangered humpback whale (*Megaptera novaeangliae*) winter in the Hawaiian Islands, and the project site lies around 1 mile inside the southernmost boundary of the Hawaiian Islands Humpback Whale National Marine Sanctuary (HIHWNMS). Humpbacks are known to frequent the entire Kona coast area in winter. The whales move

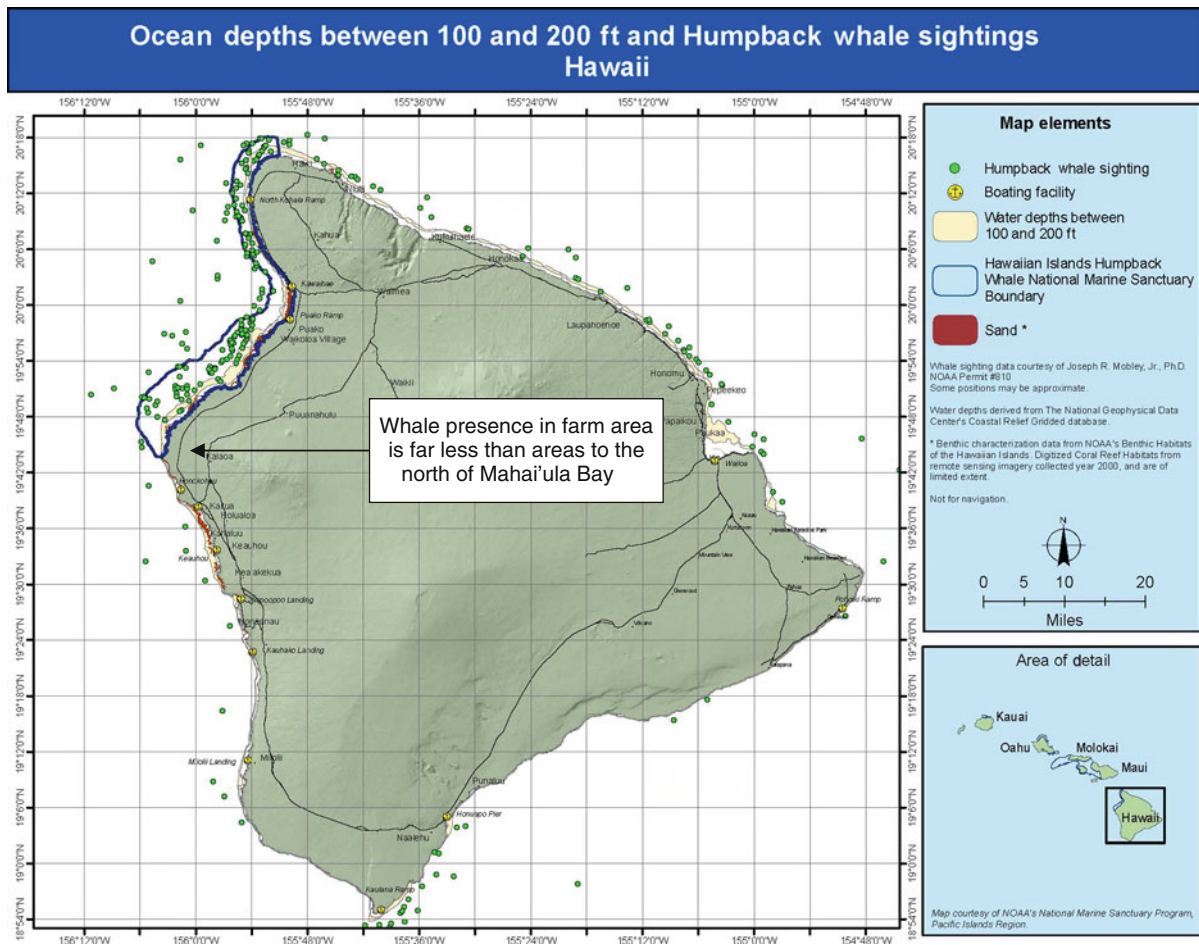
throughout the general area, usually following a longshore track (north to south, or vice versa).

Concerns about the reduction in whale habitat by the existing project were previously expressed by HIHWNMS and DLNR/DAR officials. Some concerns were also earlier expressed with the potential for entanglement of whales in the mooring lines of the net pens. A comprehensive analysis of available records of whale entanglement (NMFS Stock Assessments), a review of interactions between marine mammals and Hawaii’s fisheries [11], and details of marine mammal strandings compiled by NMFS Pacific Area Office (NMFS-PAO) shows that most whale entanglement events occur in slack net mesh (such as drift nets or fish weirs), slack vertical lines (such as crab pot or lobster pot floats), or surface lines (such as long-lining gear). Among all these observations, there is no record from any US aquaculture operation of entanglement of humpback whales, or other marine mammals, in the taut moorings or net panels of fish net pens. With heavy mooring gear, and taut lines and mesh, the potential for entanglement is considered negligible [12, 13].

Furthermore, it appears that the waters in the vicinity of Keahole Point are not as heavily frequented by the whales as are other waters of the Sanctuary, further to the north (Fig. 9). Observations from workers at the farm site suggest that the farm does not interfere with the movement of the humpback whales, beyond the immediate and obvious exclusion from the waters inside the net pens. The distance of around half-mile from the inshore side of the net pens to the shoreline offers ample room for the whales to move around the eastern end of the farm structures, without any chance for any funneling or bottleneck effects.

There is no definitive pattern of whales avoiding, or being attracted to the cages. Whales are occasionally seen within the lease area. On one instance, the farm workers witnessed a humpback on the surface inside the mooring grid array; the animal appeared to negotiate its path between the net pens and mooring lines with ease.

As part of the company’s Marine Mammal Monitoring Plan (MMMP), farm workers provide data for assessing whale abundance and patterns of movement around the farm site. The MMMP describes Federal recommendations or instructions in the unlikely event of any entanglement, and also details ongoing



Environmental Impacts of an Open Ocean Mariculture Operation in Kona, Hawaii. Figure 9
Typical Humpback whale sighting patterns around the Big Island of Hawaii

reporting requirements for any close interaction with humpback whales, or any physical interaction between the farm array and other marine mammals.

Recreational Use Impacts The farm site lies offshore from the Natural Energy Laboratory of Hawaii Authority and the Kona International Airport, and as such, has little effect on shore-based recreation. The heavily used public recreation area of Kekaha Kai State Park (Mahai'ula) lies more than 3 miles further to the north.

A survey of recreational activity in the general area, north of Keahole Point was conducted prior to the farm installation, from August to September, 2001, in conjunction with the original farm site environmental assessment [14]. The survey covered 2 months of summer conditions, which was considered the best means

of ensuring that the data represented the heaviest use of the area. The overarching finding of the survey was that the area is only used for transit: of the 150 observations made over the 61 consecutive days of the survey, only one boat was seen within the farm site – a boat transiting through the area. Most activity in the general Keahole-to-Unualoha area was recreational dive boats and commercial dive tour operations along the reef and shoreline south of Unualoha Point (directly inshore from the proposed farm site), and in Makako Bay itself.

Observations by the Kona Blue staff on the farm site suggest that this trend continues – the use of the farm lease area is merely for transit. Fishing boats now occasionally troll lines close to the central area, to try to take advantage of the aggregative effects of the net pens. There are no records of catch rates around the farm,

but anecdotal evidence indicates that catches are primarily ono (wahoo, *Acanthocybium solandri*), with infrequent catches of ahi (yellow-tuna, *Thunnus alalunga*).

Kona Blue's permit allows restricted public activities in the lease area, precluding anchoring, scuba diving, spearfishing, or swimming within the 90 acres. These limits are considered the minimum needed to protect the company's investment, to limit their liability (and retain insurance coverage), and to assure public safety. Fishing by the public from unanchored boats (trolling, or linefishing from drifting boats) is still permitted, but with the caveat that any fishing lines that become entangled in the net pen mooring lines must be left in place and cannot be retrieved by divers. The company also requests that fishermen not troll through the center of the farm site because of the potential for fishing lines to entangle divers, or for lures to hook into mooring lines or nets. Boats transiting the net pen area are also requested to observe a slow "no-wake" boat-speed to maximize safety for divers. Unguided recreational scuba diving or unauthorized commercial scuba dive tours are not permitted within the lease area because of liability, safety, and security concerns.

The loss of access to recreational activities within this relatively small area of ocean space is not considered significant. Kona Blue's ongoing observations affirm that there is virtually no fishing or other recreational use of the lease area, or the areas adjacent to the lease area, beyond trolling, which is probably enhanced by the farm's presence.

Viewplane Aesthetics Community value judgments and perceptions of how the oceans should be used largely govern the impact of the project on the community's aesthetic enjoyment of the area. In community meetings, Kona Blue generally enjoys strong support for the broad goals of the company. There is wide recognition of the severely depleted status of bottom fish species in Hawaii. The awareness of the global fisheries crisis has recently been amplified by several scientific studies, such as that of Worm et al. [15], which projected a collapse of world fish stocks by 2048, unless significant remedial changes are made to fisheries and marine ecosystem management.

The visual impact of the project is minor, compared with the adjoining properties of Kona International Airport and the aquaculture operations at the Natural Energy Laboratory of Hawaii Authority (NELHA). The major visual impact from the farm operation is from the experimental surface pens and the feed barge. There is also the additional presence of work and dive boats, and harvest boats, on some days. However, the impacts of these structures and activities are not significant, given the distance from the nearest residences, more than 3 miles away.

There is general community acceptance that the project fits in well with the overall ambience of innovative aquaculture at NELHA, and the need for Kona to develop alternative industries beyond tourism. Fisherfolk and other mariners recognize the validity of the criteria that Kona Blue has used to select this site (c.f. deeper or shallower sites), and have not expressed a strong preference for the project to be located elsewhere. Applicants for farm permits in other areas of the Kona Coast (around Kawaihae) have, on occasion, been told that their project would more appropriately be located "down near NELHA and Kona Blue."

Cultural Resources, Practices, and Mechanisms for Impact Prior to the 1801 lava flow that inundated the area, Keahole was the site of the largest fish pond in the Hawaiian islands. The Pai'ea pond (reputedly King Kamehameha's favorite pond) was approximately 3 miles long and 1/2 mile wide; canoes were used to traverse from one side to the other. The farm site is directly offshore from where Pai'ea once stood. Fish farming could therefore be considered historically and traditionally appropriate to the area.

The farm lease area is over a mile from the traditional marlin, tuna, and wahoo fishing grounds (Fig. 10). The site is too deep for free-diving or scuba diving activity but suitable for "blue-water" spearfishing. Usually, however, blue-water spearfishing is practiced close to a point or drop-off, rather than over bare sand substrate around 200 ft deep. There are no significant benthic plant or animal populations in the farm lease area, and there are virtually no benthic or pelagic fishing activities in this depth range. Kona crabs and nabeta (*Xyrichtys pavo*) are the only benthic resources that occur on sand bottom at this depth, but informants



Kona Blue's offshore fish farm site in relation to primary fishing areas. The farm site is well inside of both the 100 fathom (200 m) trolling ledge along the "grounds" offshore of Keahole Point and the 40 fathom (80 m) ono lane. Reef fishing and opelu ko'a are found well inshore of the proposed site, along the edge of the reef, in waters up to 120 ft deep (40 m). Fishing grounds for opelu at night are usually deeper than 40 fathoms (80 m)

lore, and are considered inappropriate for publication, or for sharing outside of the families or community groups who have traditionally fished these ko'a. However, in private meetings with the most knowledgeable kupuna, the locations of the traditional opelu ko'a were determined to be outside of the proposed project location [17]. Opele aggregations usually occur in water around 120 ft deep, close to reef drop-offs, and well shoreward of the farm area.

The only potentially impacted cultural resource that was cited during extensive discussions with community and kupuna (elder) groups for the original farm site was the several opelu ko'a ("holes" or schooling places for mackerel scad – *Decapterus macarellus*) that occur in the general region. The locations of these ko'a are considered to be part of traditional marine

Access to, or practice of any other customary activities has not been significantly constrained by the farm array or operations. The exclusive control over the waters (and the fish) inside the net pens is consistent with traditional and cultural practices that identified fish traps or lobster traps – and the animals therein – as the private property of the trap owner. The same principles apply here.

Global Impacts: Fish Meal and Fish Oil Usage

Fish Meal and Fish Oil Usage

Fish such as *Seriola rivoliana* (Kona Kampachi®) usually feed toward the top of the trophic chain in the wild. They therefore possess digestive systems and nutritional requirements that are adapted for feeds with high protein and lipid levels, and low levels of carbohydrates.

Fish meal and fish oil usage in fish feeds can be considered a valid use of a natural, sustainable, renewable resource, so long as the fishery from where the fish meal and fish is sourced is responsibly managed. Although stocks such as the Peruvian anchovetta fishery are sustainable in the sense that they are very well managed, they are not scalable. If mariculture is going to fulfill its potential for increasing seafood consumption to meet growing demands, then some alternative sources of proteins and oils will be required.

Kona Blue has therefore been focused on reducing the inclusion rate of fish meal and fish oil, such as Peruvian anchovies, from targeted reduction fisheries and increasing the use of agricultural oils and proteins, such as soy, canola, wheat, corn and poultry meal, and oil.

Improving Feed Conversion Efficiencies: An Evolutionary Approach

Though efficient use of fish meal and fish oil from targeted reduction fisheries is both rational and justifiable, this by no means suggests that these resources are unlimited, or that alternatives should not be searched for. If open-ocean mariculture is to develop into a food production system that can provide a significant proportion of the nutritional needs of a growing planet, then additional sources of sustainable proteins and oils for feedstuff must be found for this industry. The arc of Kona Blue's feed development strategies is perhaps instructive of

directions that open-ocean mariculture, as a global industry, might follow to achieve such scalable sustainability.

Initially, Kona Blue Water Farms fed the Kona Kampachi® with a diet that was considered “organic” by European standards. At the time, USDA did not have (and still does not have) Organic standards for aquaculture feeds. In the EU, however, Organic fish food was considered to be that which was most similar to the animal's diet in the wild. This feed, therefore comprised largely comprised fish meal and fish oil derived from Peruvian anchovies.

With the recognition of the need for more scalable feedstuff alternatives, however, Kona Blue worked with the feed vendor to develop a new diet that lowered the inclusion rate of fish meal and fish oil from Peruvian anchovies to a combined total of 50%. This diet included soybean meal, wheat gluten, canola, and other grain proteins and oils. The biological efficiency for this diet, however, was still suboptimal, with a fish-in:fish-out ratio (FIFO) of over 2:1. (i.e., an input of more than 2 lb of anchovies for each pound of Kona Kampachi® produced).

The inclusion rate of agricultural proteins in diets for marine piscivorous fish is limited by the presence of a range of “anti-nutritional factors” in the grains and less-purified meals. (Aside: although often described as carnivorous, most marine fish such as groupers, snappers, jacks, and bream, are perhaps more accurately described as “carbohydrate intolerant.” They require diets that are high in protein and lipid, and low in carbohydrate. There is no specific nutritional requirement that these fish eat meat). For this reason, soybean meal is restricted to about 20% of the diet for most marine fish. To reduce the fish meal and fish oil inclusion rate further, and to further lower the FIFO, would therefore require proteins and oils from other sources. By-products from both edible fishery processing and poultry processing were therefore included in the revised Kona Blue diet, allowing the Peruvian anchovy inclusion rate to be further reduced to 30% of the ration: 20% fish meal and 10% fish oil.

Inclusion of poultry processing by-products, however, meant that some customers, such as Whole Foods Markets (WFM), a high-end organic and natural foods retailer, would no longer carry Kona Kampachi®, even if the poultry used for the by-products was of Organic

origin. This position by WFM was out of consideration for those of their customers that were vegetarians, but still wanted to eat fish. WFM asserted that these customers would not want to eat that fish if fish had eaten a pellet that contained proteins or oils that were derived from mammals or birds. Kona Blue appealed to WFM to review their position, given the importance of reducing mankind's global footprint on the oceans, that is, the reliance of humans on natural marine resources – but as of 2010, there has been no change in this policy.

Kona Blue has recently tested two diets that completely eliminate from the Kona Kampachi® diet any fish meal and fish oil sourced from targeted reduction fisheries, and any land animal processing by-products. These innovative diets use processed by-products from sustainably managed fisheries intended for human consumption. As the trimmings from these sources would otherwise have been discarded, used as fertilizer, or burnt as fuel, the use of these fish meal and fish oil products in the Kona Kampachi® diet represents an ideal reuse of natural resources. These diets therefore would result in a zero FIFO ratio, that is, no targeted reduction fishery by-products included in the diet of the end product.

Alternative Feedstuffs for Open-Ocean Mariculture

Kona Blue is involved in testing a range of alternative feedstuffs for Kona Kampachi® diets, which also offer potential for other species of marine fish. Alternative soy products, other agricultural grain concentrates, yeast, and other single cell proteins, edible fishery by-products and – more recently, with the boom in microalgae culture for biodiesel production – defatted microalgae by-products, have all either been tested, or are under development for Kona Kampachi® feed trials.

Kona Blue has tested a range of soy-based diets, with soy protein concentrates and omega-3 oil rich strains of soybeans. These trials suggest that the inclusion rate of soy protein concentrates cannot, by itself, exceed the same 20% threshold that limits soybean meal. Above this level, growth rates and feed conversion ratios are depressed. With the inclusion of taurine in the formula, however, soy protein concentrates could replace fish meal as the source of protein up to 40% of the diet with no detriment to fish growth rates or feeding efficiencies.

There is a diverse array of edible fishery processing by-products that are available for use in aquaculture diets, and this direction offers tremendous potential for further development. The processing by-products from most wild salmon runs, for example, are woefully underutilized, and are often disposed of directly back into the rivers from which the fish are taken. Logistical and economic constraints limit the use of these trimmings, however, as the processing plants are usually small and isolated, the salmon runs are only of short duration, and storage and transport of fish meal or fish oil by-products from these villages to reduction facilities and feed mills is a challenge. Development of fish silage systems offers one potential, partial solution.

However, even the less-seasonal, larger-scale processing of farmed salmon in more centralized plants presents difficulties for the utilization of by-products. For biosecurity reasons, most fish feed plants will not run salmon-derived feed stuffs through their machinery because of the potential for contamination of feeds from viruses, bacteria, or other pathogenic vectors that may be found in the by-products. Screening for known pathogens is not an adequate solution: even though the chance could be considered very slim that some unknown pathogen may be unwittingly dispersed via extruded feed, the potential catastrophic consequences of such widespread and rapid disease dissemination are sufficient to ensure that no such chance be offered. This therefore excludes almost all large feed mills in salmon-farming regions from using salmon by-products.

Similar inefficiencies are found in the reuse of trimmings from the pollock fishery (*Theragra chalcogramma*) in the northern Pacific. This fishery primarily processes most of the catch at sea, into surimi. Trimmings from these fish constitute around 65% of the wet weight of the catch. For a fishery that has averaged around 1.3 million metric tonnes, this then represents around 850,000 tonnes of wet weight by-product annually that could be converted into fish meal and fish oil. For many years, much of this by-product was discarded back into the ocean, or the rendered fish oil was burnt in the diesel generators of the processing vessels. Some 8 million gallons of fish oil in Alaska is largely disposed of as biodiesel [18]. More recently, some proportion of these trimmings have been used to make a high-quality white fish meal that is largely exported to Asia, where it is valued in feeds for

farmed eels. However, the proportion of by-product that is reused or recycled is not reported. Again, economics and logistics conspire against development of a rational supplement to targeted reduction fisheries. The increasing prices of fish meal and fish oil, however, driven by the growing demand for animal feeds from developing economies (notably China and India) may be a greater incentive to resolve these constraints. Edible fishery by-products may yet play a significant role in aquaculture feedstuff sourcing.

There is ample evidence that some or all of these innovative feedstuffs could help to reduce the demand for fish meal and fish oil from clupeids in the medium- to long-term.

Is Aquaculture “Fishing down the Food Chain”?

Much interest has been recently focused on the problem of “fishing down the food chain,” beginning with Pauly et al. [19], and Taylor et al. [20]. This is the trend over time for commercial fisheries – driven by serial stock depletion – to shift their target species to those lower on the trophic pyramid. Fishermen first start out exploiting the high-value, top-end predators, then move on to mid-level predators, and then down toward herbivores and detritivores – what was previously considered bycatch. Fisheries generally start out targeting the larger, sweeter-tasting species – tunas, snappers, groupers, and such. As these become increasingly scarce, fishermen apply greater fishing power, and fish longer and deeper, retaining or targeting what was previously considered “trash.” The argument portends that at some stage, the food web is reduced to an ocean full of jellyfish. “Fishing down the food chain” is a condemnation of the inherent unsustainability of most commercial fisheries management – or rather, mismanagement.

Fish farming has somehow been implicated in this practice on the basis of farmed fish being fed pellets that are partly made up of fish meal and fish oil derived from anchovies, menhaden, sardines, or the like. These fish (collectively, the clupeiforms) usually form the first step in the ocean food chain beyond primary production. Some scientists and anti-aquaculture advocates misconstrue or deliberately misinterpret the complexities of ecological and economic cause-and-effect, and represent the use of clupeiforms as feed for farmed fish

as wanton. This has been led by respected institutions such as the Monterey Bay Aquarium [21, 22], but has also spilled into mainstream media, such as the *NY Times* [23], *Conservation Magazine’s* article on “10 Solutions to Save the Ocean” [24], *The Ecologist* [25], and *The Economist* [26]. The notion that aquaculture is guilty of “fishing down the food chain” is now lodged within the public consciousness.

The bottom of the food chain, however, is where fishing should preferentially be done. It makes far more sense to use herbivores or planktivores from the base of the trophic pyramid as either human food or feed for farmed fish, than to be targeting top-end predators. This makes economic sense, but it also makes sense from other perspectives: it is better for the ocean’s ecosystems, it is better from the viewpoint of bioenergetics transfer through the trophic pyramid, it is better for consumer health, and it makes for better fisheries management.

The economics are simple: Peruvian anchovies and menhaden are not highly valued in the market, so they are cheap. Maybe this will change in time, and prices for anchovies and sardines will increase, as more people develop a taste for oily baitfish, but it is more likely that most consumers will still prefer larger piscivorous marine fish as sashimi or fillets.

The ocean’s ecosystem offers several reasons why the bottom of the trophic pyramid is a better place for humans to extract nutrition from the sea. It is, most simply, a matter of mass and mathematics. Herbivorous fish are more abundant, with greater biomass. Catching 1,000 t of Peruvian anchovies has little impact on the 6 million ton spawning biomass (around 0.025%). By contrast, 1,000 t of tuna represents around 10% of the Bluefin Tuna spawning stock in the Western Atlantic, the biomass of which is currently estimated at less than 10,000 t [27].

Moreover, Clupeiforms are classic “r-selected” species, with their smaller body size, faster maturing, and shorter life spans [28]. They are highly opportunistic: a decrease in population size in Peruvian anchovies often results in increased recruitment from the next spawning. From an ecological perspective, these species are precisely where fishing effort should be targeted, not the larger, more vulnerable, slower-growing “K-selected” species at the top of the food chain. In agricultural terms, most of the crops that humans raise

are strongly “r-selected” – wheat, corn, barley, rice, while targeting a “K-selected” species in agriculture might be the equivalent of chopping down oak trees to eat the acorns.

Herbivorous clupeiforms also grow and reproduce faster. Menhaden stock resilience to fishing pressure is “high” [29], with a population doubling time of only 15 months. Northern Bluefin Tuna, by contrast, have “low” stock resilience, and a minimum population doubling time of 4.5–14 years [30]. Therefore, if half the menhaden were harvested, it would take 15 months for the stock to recover. However, if half the tuna population was taken, it would take, at a minimum, between 4.5 and 14 years to recover. Southern bluefin tuna also do not begin to spawn until they are perhaps 11 years old, and may live to “at least 40 years of age” [30]. However, Peruvian anchovies are sexually mature within 1 year, and only live for around 3 years. The 3-year old anchovies then die and fall to the ocean floor.

The public health imperative should also provide impetus to source fish meal and fish oil from lower down the food chain. Menhaden and anchovies filter algae and zooplankton directly from the water. They are therefore high in heart-healthy omega-3 oils, yet low in the persistent organic pollutants, such as mercury and PCBs. These pollutants, however, are concentrated as they move further up the food chain. It is primarily top-level predators – sharks and tuna – that are on FDA advisories for pregnant and nursing mothers, and children. By contrast, an aquaculture species that can achieve a feed conversion efficiency of close to 1:1 (FIFO, or Fish In : Fish Out) contain essentially the same contaminant loading as the clupeiforms at the base of the food chain.

Clupeiform fisheries are also more readily managed, with relatively simple stock dynamics and ecosystem interactions. The major inputs to clupeiform stocks are the spawning biomass and primary productivity, which is usually driven by the strength of the nutrient-rich upwelling. Most of the fisheries occur within the EEZ of a single nation, where there are direct incentives for sound management and enforcement, and where access can be regulated. Tuna and swordfish, by contrast, are highly migratory species. Donut-holes of high-seas waters, beyond any country’s 200-mile zone, provide opportunities for distant-water fishing nations to concentrate their boats and effort. Attempts

at managing tuna stocks are typified by the International Commission for the Conservation of Atlantic Tunas (ICCAT), which has 46 members, and almost no enforcement capabilities. And while Hawaii’s long-line fishery targeting big-eye tuna may be very well managed, for example, heavily subsidized European or Asian purse-seiner fleets target the juveniles of the same stock in the South-Western Pacific.

Moreover, the carbon footprint of clupeiform fisheries is minimal. These fish are usually taken by purse-seiners, working close to the coast, encircling schools containing hundreds of tons at a time. The carbon footprint for species higher on the foodchain is much higher, with fish being caught by diesel-powered trawlers or trollers, or – for bluefin tuna and swordfish – by harpooning the fish, one at a time.

Most importantly, however, it is far better from a bioenergetic perspective to target fish closer to the bottom of the food chain. Applying the 10% trophic transfer rule means that the 1 lb of wild tuna sashimi on a consumer’s plate needed to eat 10 lb of anchovies – or its equivalent in fish meal and fish oil. Or maybe, if there were two steps in the trophic pyramid, each pound of wild tuna required 100 lb of anchovies to first be converted into 10 lb of mackerel.

Aquaculture, by contrast, is always a single step – clupeiforms-to-crop. But aquaculture can also use alternative agricultural proteins and oils, such as corn and wheat gluten, soy proteins and oils, canola, and other animal processing by-products. These other proteins and oils reduce the fish meal and fish oil inputs to the extent that some of the purported “carnivores” can thrive on a diet that is around 20% fish meal and fish oil. On the “sustainability quotient” – the number of pounds of fish-in to produce 1 lb of fish-out (the FIFO) can then attain the perfectly efficient goal of parity, or 1:1 (i.e., every pound of Peruvian anchovies in the farmed fish diet produces 1 lb of product). The result is efficient conversion of a low-value anchovy into a high-value marine fish, without disrupting the fragile top of the trophic pyramid.

Larger, wild fish are more bioenergetically wanton. Wild fish lose energy through inefficient digestion, in hunting prey, trying to avoid predation, spawning, and succumbing to natural mortality. As wild fish grow larger, they also become increasingly inefficient – a greater proportion of energy is needed to maintain the animal’s

metabolism. Any bycatch will compound these inefficiencies further. The global bycatch ratio is around 0.28 lb of discard for every pound of target species.

Earlier estimates [20] suggested that farmed fish might be more efficient than wild fish, based on a single trophic step, by a factor ranging from 2 to 5. Combining the life cycle inefficiencies, trophic inefficiencies, and bycatch inefficiencies of wild fish, however, means that farmed fish may be more efficient than wild fish by a factor of around 60 (Table 1, and [31]).

This reasoning does not advocate for greater fishing effort on anchovies. To the contrary – caution is called for. While most of these stocks are sustainably managed at current levels, they could not withstand any greater pressure. These clupeiform stocks should continue to be very closely monitored and highly regulated. Large marine protected areas should also be established to allow some clupeiform-based ecosystems to flourish in their natural state (rather than attempting ecosystem based management). But it is imperative that the

Environmental Impacts of an Open Ocean Mariculture Operation in Kona, Hawaii. Table 1 Relative ecological efficiencies of farmed and Wild-Caught fish. The table shows the compounded cost in terms of anchovy-equivalents for farmed and wild-caught fish. Low-end estimates and high-end estimates are provided for each type of fish and compared cross-ways to obtain a lowest relative rate and highest relative rate

	Farmed fish		Wild-Caught fish		Global mean
	Low-end estimate	High-end estimate	Low-end estimate	High-end estimate	Ratio of wild to farmed
Life cycle efficiency ^a	1	1	3	10	6
Trophic transfer efficiency ^b	1	8	10	100	7.3 ^c
“Bycatch” efficiency	1	1	1 ⁽⁴⁾	11 ^e	1.3 ^d
Compounded “cost”	1	8	30	11,000	57

The lowest relative rate extrapolated from this table is that the least-sustainably farmed fish are around 4× more ecologically efficient than the most sustainably harvested wild fish (i.e., 30:8). The highest relative rate is that the most sustainably farmed fish could be 11,000× more ecologically efficient than the least-sustainably harvested wild fish (i.e., 11,000:1). The Global Mean of wild fish efficiency to farmed fish efficiency is around 57×

^aThere are no published estimates of the relative life cycle efficiencies of farmed versus wild fish. However, fish that reach reproductive age in captivity can see Feed Conversion Ratios increase by factors of 5 or 10 over juvenile and sub-adult fish. Natural mortality and the nutritional cost of maintenance of basal metabolic processes during periods of food deprivation also increase the “Economic” Feed Conversion Ratio for wild fish populations

^bIn 1997, food conversion efficiencies (FCE) for farmed marine fish and farmed salmon were around 5:1 and 3:1, respectively (Naylor et al. 2000). By 2010, however, FCEs are projected to reach 1.5:1 for farmed marine fish, and as low as 1.2:1 for farmed salmon (Tacon 2005). Kona Blue has been able to culture Kona Kampachi® on a diet that equates to a 1:1 ratio of wet-fish-in to wet-fish-out. However, if a less-sustainably farmed fish is fed a pellet high in fish meal and fish oil (say, to meet the Scottish Soils Association’s Organic standards, with around 80% fish meal and fish oil), this diet could equate to around 4 lb of wet anchovy-equivalents for every 1 lb of dry pellet (a wet-fish to fish-meal ratio of 5:1 is considered standard). On this diet, most commercially farmed species might have food conversion ratios of around 2:1 (dry pellet to wet fish), implying an FCE of 8 lb of wet-fish-in for every 1 lb of wet-fish-out

^cTacon’s (ibid) estimate of FCEs for farmed salmon and farmed marine fish might be conservatively pooled at, say, 1.5:1, that is, 1.5 lb of anchovy-equivalents for every pound of farmed fish produced worldwide. There is a differential of around 1.1 trophic levels between global fishery landings (with a mean trophic level of around 3.3) and the Peruvian anchovetta fishery (with a trophic level of around 2.2: Pauly et al, [19]). At a presumed 10% biomass transfer efficiency up each trophic level, this implies 11 lb of anchovy-equivalents to produce a pound of harvested wild fish. The median ratio of wild to farmed trophic transfer efficiencies can therefore be estimated at 11:1.5, or 7.3:1 overall

^dHarrington et al. (2005), report a “nationwide discard-to-landings ratio of 0.28” (i.e., for 3.7 million tons landed, some 1.06 million tons were discarded). However, for highly selective fishing methods, such as harpooning, bycatch is effectively zero, as for farmed fish

^eFor finfish, the ratio of bycatch to target fish (in the Northern Pacific) can be as high as 11:1 because the bycatch is either too young, out of season, or the vessel has no permit to keep it (Alverson 1998)

fisheries at the base of the food chain be better managed, and environmentally sound aquaculture endorsed so that pressure can be taken safely off the top of the food chain.

Fishing at the bottom of the food chain should therefore be encouraged preferentially over any other kind of fishing. This is not a function of recent overfishing: even 100 years ago, this principle would have still held true. People always *should have been* fishing at the bottom of the food chain. To continue to accuse aquaculture as being part of the problem of “fishing down the food chain” is therefore disingenuous. Aquaculture is an important part of the solution to the feeding of the growing humanity. To assert otherwise confuses the consumer and discourages the policy shifts needed toward more sustainable aquaculture, and healthier oceans.

Future Directions

The Challenge

Open-ocean mariculture must expand, but it must expand in an environmentally responsible manner. While opponents may cite the precautionary principle as reason to not move forward, or to do so only cautiously, with experimental site permits or restrictive legislation, there is an imperative for action. Extractive pressures cannot be increased on already depleted wild fishery resources, and the public health costs of limited seafood consumption cannot be accepted any more. These two trends must be turned around by finding alternative sources of healthful seafood. Open-ocean mariculture is the only practical means of achieving this. To urge inaction, then, is in effect advocating for either a greater fishing pressure on wild stocks, or the increasing human mortality and suffering from heart disease and stroke. If the consequences of inaction are so inevitable and so severe, and the potential consequences of action are only slight and temporary, then the precautionary principle insists that action must be taken. Therefore, the requisite regulatory framework for industry growth must expediently be set in place and the needed technologies and support industries developed to proceed forward.

The single greatest challenge will remain that of effecting a change in the mind-set of the environmental community to the extent that there is broad acceptance

of the need for expansion of open-ocean mariculture as a conservation tool, that is, as an alternative to targeting larger wild fish. Until this is accomplished, then, this industry will continue to be smeared with irrational fears, community prejudice and consumer bias, and the promise it offers will be left woefully unfulfilled.

Further! Deeper!

The Kona Blue operations to date have produced annual harvests of around 500 T, and have demonstrated clearly that this can be achieved without any significant impact on ocean ecosystems or resources. The key now is to ensure that these impacts remain insignificant, as the industry scales. Expansion opportunities, in terms of broader acceptance and greater access, will be conditional on future farm sites being located further offshore and in increasingly deeper waters. The industry therefore needs to develop the technologies to support these trends and to ensure that carrying capacities and prudent biosecurity practices are adhered to as the industry grows.

There are economic and environmental drivers for larger-scale open-ocean mariculture operations in deeper water, further offshore. As the automation systems for net pen management become increasingly sophisticated [32], sea state becomes less of an impediment to more remote, more exposed sites. The major disincentive then becomes travel time to and from a site. Beyond about a 20 mile distance, it is increasingly difficult to maintain a farm site with day-workers, and some on-site residence becomes necessary. This, then, argues for further scaling: once a farm site is manned 24 h, with staff changes twice-weekly or weekly and periodic delivery of feed and other supplies, there are strong commercial reasons for the operation to grow in scale to support the inherent fixed costs.

Larger scale net pens present some challenges in the open ocean, particularly with fish handling, but there are advantages as well. Larger pens are more cost-effective in terms of cost per unit of volume (i.e., cost per cubic meter) and in terms of managing the stock. A farmer manages a net pen, rather than a fish: a net pen of 3,000 m³ and 50,000 fish requires roughly the same level of management as a net pen of 24,000 m³ and 400,000 fish. Similarly, where there is a significant

cost to the netting material, or cost to maintaining the netting, then the lower surface area to volume ratios of larger net pens are increasingly attractive. Larger net pens also require larger hatcheries, to produce sufficient fish in one cohort to stock the pen, and this results in greater hatchery efficiencies.

New netting materials under development could revolutionize the net pen designs. Kikkonet™ appears to offer significant advantages over multi-strand nylon or cotton webbing, with greater rigidity and predator-breach resistance reducing the risk of escapes and easier cleaning capabilities with the single monofilament material. Similar advantages may be offered by new brass alloy materials under development, which almost completely eliminate biofouling and the attendant fish health concerns, as well as offering improved resistance to breaches from predators or mechanical tearing. Both Kikkonet™ and brass webbing, however, require rigid frames in open-ocean environments: constant wave and current movement can quickly destroy netting material if the net pen frame allows any excess movement. Each of these materials represents a significantly higher capital cost than nylon or cotton netting, but there are tremendous operational savings to be gained from reduced risk of failure and reduced labor for net changing, net washing, antifoulant dipping, and net loft work.

Although much discussion has focused on the advantages of surface and submerged net pens and the relative operational efficiencies of each, this may be the single most important deciding factor for pen design: rigid-framed nets such as the submersible Sea Station® and Aquapod® can support more robust, rigid netting materials. Flexible framed nets such as PolarCirkel®, or those that rely on gravity to maintain net shape, such as Wavemaster® can only support pliable nets of nylon and cotton in the higher wave and current conditions offshore. Where sea conditions or predator prevalence demand the more robust, rigid mesh forms of brass or Kikkonet, then submersible rigid-framed pens would appear to have operational advantages.

Letting Go

The potential for untethered “trans-ocean drifter net pen” fish farms has, up to now, been the stuff of science

fiction [3]. However, the concept of untethered pens offers the potential for far larger scaling of operations, with almost negligible potential for any environmental impact; a net pen in very deep water is a nonpoint source for effluent impacts on water quality and substrate, and there is almost no potential for interaction with wild fish stocks (apart from aggregative effects). There are also advantages for fish health management on the farm – a drifter pen is essentially perpetually fallow. Drifter pens that ride regional ocean eddies or powered pens (either towed by a surface vessel, or where the fish containment is integrated into the vessel hull) are already moving from concept to prototype; within 5–10 years, some form of untethered net pen may become a commercial reality. Several examples are worth noting:

In June, 2008, a trial in Puerto Rico tested a submerged, self-propelled, 3,200 m³. Aquapod as a demonstration of untethered aquaculture operations. The system was able to achieve 30 cm/s forward speed by using two electrically driven 2.5 m diameter propellers powered by a diesel generator set onboard a small towed vessel. The Aquapod was able to hold position in 25 cm/s currents and was easily maneuverable. The concept of remote-controlled or autonomous operation was judged feasible using cage-mounted sensors and buoy-mounted GPS and communications.

In October 2009, the State of Hawaii approved a tuna farm permit for a 247 acre site offshore of the Kohala Coast (north of the Kona Blue farm site) in water over 1,320 ft deep, for up to 12 untethered Oceanspheres. With no anchors, it is proposed that these fish farming platforms will be held in position by self-propulsive forces powered by a hybrid Solar Ocean Thermal Energy Generator. Hawaii Oceanic Technology, Inc. plans to deploy the first of these net pens in 2011.

A regional drifter pen concept under development by Kona Blue – christened the *Veleva* system – is projected to remain within a prescribed geographical area without anchoring, by riding the eddies in the lee of oceanic islands such as Hawaii, or by riding the gyres found in semi-enclosed seas, such as the Sea of Cortez or the Gulf of Mexico. The advantages of this concept are that it then enables the operation to consistently

stock, feed, tend, and harvest the fish in the pen, as well as changing crew and other maintenance tasks. By being untethered, the *Velella* can be located in very deep water, beyond normal mooring depths, and any potential environmental impacts are reduced to de minimis levels.

As untethered net pens do not occupy a fixed geographical space, they do not require a lease but rather are more properly regulated as non-powered vessels. Where there is no point source effluent from a motile pen, the monitoring and regulatory requirements may also be those applied to vessels, rather than fixed farm locations. Projects such as *Velella* might also pass through or be deliberately sited in international waters, where no permits are presently required and where no regulatory framework is currently in place. This regulatory vacuum should be remedied, to both govern its growth but also encourage the development of this technology.

IMTA

Much interest is often focused on Integrated Multi-trophic Aquaculture (IMTA) as a means of improving the sustainability of open-ocean fish farming, by mitigating effluent impacts and more fully utilizing the nutritional inputs. In nearshore systems, the siting of filter-feeding bivalves around fish farm net pens in IMTA systems has been shown to reduce particulates in the water column. Similarly, siting macroalgal culture systems downcurrent of the fish farm can also reduce dissolved organics. These polyculture systems also offer secondary products (such as edible seaweeds and mussels) that are themselves marketable.

In open-ocean mariculture systems, however, the applicability and utility of IMTA is somewhat dubious. The priority for siting open-ocean mariculture operations should be to strive for greater water clarity, depth, and current movement to optimize fish health and minimize any potential impacts on the effluent. These criteria are contrary to the desired characteristics of an IMTA operation, which presumes some entrainment of waters after it passes through the fish pen, with enrichment of the effluent providing benefits to the filter feeders and phototrophs. If open-ocean mariculture

operations do indeed have no measureable impact on water quality, then presumably there is no benefit to be gained from co-siting the fish with filter feeders and plants.

Furthermore, the capital and maintenance costs for the secondary production systems would be prohibitive. Open-ocean mariculture, in the foreseeable future, will need to focus on high-value species such as tuna, yellowtails, snappers, and groupers to provide a reasonable return. Only with the development of far greater operational efficiencies, in later years, will lower-value finfish species become economically viable. The commercial returns from deploying and maintaining even lower-valued macroalgae and bivalve culture lines around an offshore site would be even further deferred.

In addition, the co-siting of mussels and seaweed, say, around an open-ocean site represents engineering and biological risks to the fish farm operation. The additional drag from the long-lines would add tremendous strain on the mooring system for the net pens. If moored separately, there would still be concerns that the secondary producer moorings could fail under load and entangle with the fish's net pen or moorings. The resting stages of pathogens or parasites might also be harbored in the cultured species or ecosystem that surrounds the long-lines, resulting in a proliferative feedback loop and elevated levels of these pests in the cultured fish.

The overarching goal of open-ocean mariculture is growing high-value species in near-pristine waters, with little potential for their interaction with surrounding biologically active substrates. The notion of burdening an offshore operation with lower-value production systems is antithetical to this goal. The fish carrying capacity should allow the natural assimilative capacities of the surrounding ecosystem to fully absorb the farm's ecological footprint. The nutritional by-products, that is, the nutrients in the effluent from open-ocean mariculture systems can still promote productivity on a broader scale, especially in oligotrophic tropical waters. However, a deeper ecological perspective would suggest that there is no need for man to capture this productivity for commercial gain. The oceans can use these inputs in other ways; they are not necessarily "lost."

Bibliography

Primary Literature

1. FAO (2009) The state of world fisheries and aquaculture. FAO, Rome. www.fao.org/docrep/011/i0250e/i0250e00.HTM
2. Clarke AC (1957) The deep range. Harcourt, New York, p 238
3. Mann CC (2004) The bluewater revolution. Wired Magazine, Issue 12.05 – May 2004
4. Environmental monitoring data for Kona Blue. www.kona-blue.com/emonitoring.php
5. See 2003 Final EA. www.kona-blue.com/permitapplication.php
6. Tagawa AW, Tam CKM (2006) Hawaii's Ulua and Papio tagging project 2000 to 2004. DAR Technical Report 06-01. DLNR, Honolulu, p 75
7. Meyer C (Personal communication, 2/17/2010)
8. Balazs GH (1980) Synopsis of biological data on the green turtle in the Hawaiian Islands. U.S. Dep. Commer., NOAA Tech. Memo. NOAA-TM-NMFS-SWFC-7, p 141
9. Viezbicke J (Personal communication)
10. 2009 Draft EA Appendix 2: Marine Mammal Report from Kona Blue to NOAA, dated 11/26/08, www.kona-blue.com/download/REVISED_DraftEnvironmentalAssessment2.pdf
11. Nitta ET, Henderson JR (1993) A review of interactions between Hawaii's fisheries and protected species. Mar Fish Rev 55(2):83–92
12. Celikkol B (1999) Biological assessment of the University of New Hampshire open ocean aquaculture demonstration project finfish component. Prepared for U.S. Army Corps of Engineers, New England Division, Concord, p 60, plus appendices
13. Wursig B, Gailey GA (2002) Marine mammals and aquaculture: conflicts and potential resolutions. In: Stickney RR, McVey JP (eds) Responsible marine aquaculture. CAB International, Wellington, pp 45–59
14. See Appendix III, Black Pearls, Inc., Final Environmental Assessment, 2003, <http://www.kona-blue.com/permitapplication.php>
15. Worm B, Barbier EB, Beaumont N, EmmetDuffy J, Folke C, Halpern BS, Jackson JBC, Lotze HK, Micheli F, Palumbi SR, Sala E, Sekoe KA, Stachowicz JJ, Watson R (2006) Impacts of biodiversity loss on ocean ecosystem services. Science 314(5800):787–790
16. Punihaole R (2003) Personal communication
17. Maly K, in litt, 8/26/02
18. Alaska Energy Authority, undated brochure
19. Pauly D, Christensen V, Dalsgaard J, Froese R, Torres F Jr (1998) Fishing down marine food webs. Science 279:860–863
20. Taylor R, Goldburg RJ, Primavera JH, Kautsky N, Beveridge MCM, Clay J, Folke C, Lubchenco J, Mooney H, Troell M (2000) Effect of aquaculture on world fish supplies. Nature 405:1017–1024
21. O'Neill B (2006) Seafood Watch® Seafood Report: U.S. Farmed Rainbow Trout, *Oncorhynchus mykiss*. Final Report. June 23, 2006. Monterey Bay Aquarium, p 26. www.montereybayaquarium.org/cr/cr_seafoodwatch/content/media/MBA_SeafoodWatch_FarmedTroutReport.pdf
22. Miranda IT, Peet C (2008) Seafood Watch® Seafood Report: Farmed Yellowtail, *Seriola* spp. Final Report. 22 Oct 2008, Monterey Bay Aquarium, p 73. www.montereybayaquarium.org/cr/cr_seafoodwatch/content/media/MBA_SeafoodWatch_FarmedYellowtailReport.pdf
23. Bittman M (2008) A seafood snob ponders the future of fish. The New York Times, 16 Nov 2008, p. WK1. <http://www.nytimes.com/2008/11/16/weekinreview/16bittman.html>
24. Simpson S (ed) (2007) 10 solutions to save the ocean. Conservation Magazine 8(3)
25. Wasley A, Wickens J (2008) Fishy business. The Ecologist, Dec 2008–Jan 2009, pp 24–29
26. The Economist (2009) Come, friends, and plough the sea – but make sure the farmed fish are veggies. The Economist: A Special Report on the Sea. 3 Jan 2009, p 11
27. ICCAT (2008) Report of the Standing Committee on Research and Statistics (SCRS); Madrid, Spain, September 29 to October 3, 2008. BFTW-Figure 2., p 89. International Commission for the Conservation of Atlantic Tuna, Madrid, Spain. http://www.iccat.int/Documents/Meetings/Docs/2008_SCRS_ENG.pdf
28. Pianka ER (1970) On r and K selection. Am Nat 104:592–597
29. Fishbase (2009a) *Brevoortia patronus*, Gulf menhaden. <http://www.fishbase.org/Summary/SpeciesSummary.php?id=1589> (1/30/2009)
30. Fishbase (2009b) *Thunnus thynnus*, Northern bluefin tuna. <http://www.fishbase.org/Summary/SpeciesSummary.php?id=147> (1/30/2009)
31. Pauly D, Christensen V (1995) Primary production required to sustain global fisheries. Nature 374:255–257
32. Langan R (2010) Technology needs for improved operational efficiency of open Ocean Cage Culture. J Marine Tech Soc 44(3):47–55

Books and Reviews

- Bridger CJ, Costa-Pierce BP (2003) Open ocean aquaculture – from research to commercial reality. The World Aquaculture Society, Baton Rouge, p 351
- ICES (2006) Report of the Working Group on Environmental Interactions of Mariculture (WGEIM). Narrangansett, Rhode Island, U.S.A. ICES CM 2006/MCC:03, p 195
- IUCN (2004) Mediterranean marine aquaculture and environment; Identification of issues. IUCN Center for Mediterranean Cooperation, p 34
- Halwart M, Soto D, Arthur JR (eds) (2007) Cage aquaculture. Regional reviews and global overview. FAO Fisheries Technical Paper No. 498. UN-FAO, Rome, p 254
- Lee CS, O'Bryen PJ (eds) (2007) Open ocean aquaculture – moving forward. Oceanic Institute, Waimanalo, p 78
- Marine Aquaculture Task Force (2007) Sustainable marine aquaculture: fulfilling the promise; managing the risks. Woods Hole Oceanographic Institute, Woods Hole, p 128
- NOAA (2008) Offshore aquaculture in the United States: economic considerations, implications and opportunities. NOAA, Silver Spring, p 272. <http://aquaculture.noaa.gov>

Environmental Radioactivity and Ecotoxicology of Radioactive Substances, Introduction to

GLEN A. BIRD

EcoMetrix Incorporated, Mississauga, ON, Canada

Radioecology is a combination of radiobiology and ecology that involves the study of the behavior of radionuclides in the environment and their effects on organisms. Radioecology is multidisciplinary involving studies of pathway analysis, speciation, bioaccumulation, food chain/food web transfer, radionuclide cycling, biological effects (radiobiology), modeling, prediction, environmental protection, risk assessment, and remediation, and may involve either ecosystem components or whole ecosystems. The goal of radioecology is to better understand, predict, and prevent environmental impacts due to radiation. This frequently involves the use of quantitative predictive models to assess the movement of radionuclides in the environment, and their transfer to biota where effects may be at the genetic, tissue/organ, individual, population, or community level.

The articles comprising this section describe the sources of radiation in the environment, the behavior of radionuclides in environmental media, their transfer to biota and potential radiation effects on biota. The basis and content of each of the 15 articles in this section are described below.

Radionuclides are ubiquitous in the environment and may be naturally occurring or the product of nuclear technology. Natural radiation consists of primordial radionuclides present since the formation of the earth and cosmogenic radionuclides. Primordial radionuclides include the ^{238}U , ^{232}Th , and ^{235}U decay chain radionuclides and ^{40}K which forms a small percentage of stable K (0.017%) present in all living organisms. Naturally occurring radioactive material is called NORM, whereas technologically enhanced naturally occurring radioactive material is called TENORM and includes mine/mill tailings, fertilizer, pulp and paper, and petroleum. Other sources of radiation are historic weapons testing, nuclear power plant operations, accidents (Chernobyl, Kyshtym), medical and industrial

usage, and waste disposal ([► Radiation in the Environment, Sources of](#)).

In the nuclear fuel cycle, the mining and milling of uranium, reprocessing of spent fuel, and waste management represent the largest releases of radionuclides to the environment. Uranium is a common contaminant released from these activities. Uranium is more chemotoxic than radiotoxic and can be harmful at low concentrations especially to phytoplankton and zooplankton. Absorption of uranium in the digestive track of mammals is low and uranium does not biomagnify up the food chain. In vertebrates, damage is primarily to the kidney ([► Uranium in the Environment: Behavior and Toxicity](#)).

Tritium is a radioactive hydrogen atom that may be present as tritiated water (HTO), tritiated gas (HT), or organically bound tritium (OBT). Tritium exchanges readily with hydrogen atoms in media including water, so tritium moves rapidly through the environment by evaporation, transpiration, diffusion and isotope exchange, and enters biota and humans via inhalation and ingestion. Photosynthesis is the most important process converting tritium into OBT in complex molecules, increasing its retention in organisms. Dynamic models are available that account for isotopic discrimination and involve studies of atmospheric releases, model validation and uncertainty, and are process oriented ([► Tritium in the Environment](#)). Because tritium is a weak beta emitter, it is an internal hazard when inhaled, ingested, or absorbed into the body. Biological modeling of tritium in the body involves the use of dose models for inhalation and ingestion of tritiated water based on specific activity concepts to assess potential effects such as teratogenic, reproductive, and carcinogenic effects ([► Tritium, Health Effects and Dosimetry](#)).

Within the terrestrial system the soil is the ultimate repository for radioactivity. Radionuclides are assimilated into soil organisms (microbes, invertebrates, and plant roots) or immobilized in abiotic components. Important properties of radionuclides are their biological half-life and ecological half-life. Chemical properties of radionuclides control their behavior and transport. Daughter products may have significantly different chemical properties from their parent. The soil hydrology is important in controlling the transport

and the buildup of radionuclides. Particle surface area to volume ratio, cation exchange capacity, metal oxide and hydroxides, organic matter, cation exchange, precipitation/dissolution, surface complexation, redox reactions, plant uptake, transfer factors, solute partitioning, dispersal coefficients, and retardation all interact to influence the behavior and effect of radionuclides in the soil environment ([► Radionuclide Fate and Transport in Terrestrial Environments](#)).

Soil erosion robs agricultural land of its productivity, deposits sediment in downstream areas and impacts water quality. Distinctive behavior of fallout radionuclides (^{137}Cs , ^{210}Pb , ^7Be) allow the quantification of sediment mobilization and deposition rates, tracing movement of fine sediment and associated contaminants through the landscape, discriminating the sources of sediment, and how this changes with change in land use/management practices ([► Fallout Radionuclides and the Study of Erosion and Sedimentation](#)).

In the arctic, lichens are the basis of the lichen/caribou/human (wolf) food chain and are important in the transfer of radionuclides to top predators. Lichens have a large surface area and no roots but scavenge nutrients/radionuclides from the air. Caribou and reindeer (*Rangifer tarandus*) are the only mammalian species that can subsist on a diet of lichen over the winter. Caribou populations are presently on the decline possibly because contaminants such as radiation may interact with other environmental stressors and weaken the overall ability of the animals to survive. Long-lived-artificial radionuclides (^{137}Cs and ^{90}Sr) and natural radionuclides (^{238}U and ^{232}Th decay chain radionuclides) accumulate in caribou. Following the Chernobyl accident high levels of radiation were found in Swedish reindeer, and chromosomal aberrations were observed in Norwegian reindeer calves ([► Radiation Effects on Caribou and Reindeer](#)).

To assess the contribution of foodstuffs to the internal doses received by humans and wildlife, models rely on transfer parameters which describe the empirical ratios of activity between two compartments. Transfer parameters such as soil to plant concentration ratios are lumped parameters that account for the various processes and interactions that result in radionuclide concentrations in these media: speciation, oxidation, reduction, transformations and interactions with soil

components, clay and organic matter content, stable element concentration, pH, land management, and soil solutions. Transfer coefficients to farm animals consider metabolic rate, milk production rate, soil intake, stable element intake, transfer of radioiodine to milk, partition coefficients, and allometric relations ([► Radiation Assessment, Use of Transfer Parameters](#)).

Past practices and accidents have left radiation legacies. Remediation of these contaminated sites is receiving considerable attention. Many factors govern the effectiveness and cost of countermeasures, most of which are site specific. Environmental restoration is a complicated multidimensional task which needs to take into account radiological, ecological, and social-economic issues within a holistic approach. The ICRP principles of protection: justification, optimization, and dose risk limitation cost benefit analysis should be applied. Practitioners need to identify suitable, effective individual procedures, remediation technologies, communication strategies, and decision and management options. Spatial factors such as size of affected area, direct protection of the population, restriction of food, percent reduction in dose, and effectiveness must be considered ([► Environmental Countermeasures and Restoration](#)).

The behavior of radionuclides in freshwater ecosystems depends on the chemical characteristics of the radionuclide, the nature of the waterbody, and the physical state (particulate, dissolved, gaseous, colloidal) of the radionuclide. The chemical composition of the water (pH, dissolved organic content, alkalinity, hardness), partition coefficient, precipitation, uptake by phytoplankton, sedimentation, transfer to sediment, and sediment focusing interact to govern the behavior of radionuclides. Transfer to sediment depends on whether the radionuclide behaves as a conservative (nonreactive – remains dissolved in the water column,) or nonconservative (reactive – has a strong affinity for suspended particles and sediment) element, sediment type, organic and clay content, water temperature, hydrological flushing rate, lake/river morphology, redox conditions, decomposition, food web transfer, feeding behavior, and trophic level of organisms (See section on “Aquatic Ecosystems” in [► Uranium in the Environment: Behavior and Toxicity](#)).

The release of radionuclides to the marine environment such as following the Fukushima accident in Japan

in April 2011 will result in interactions with individual organisms and food webs. The speciation of radionuclides in seawater differs from that in freshwater and affects the mobility of radionuclides; their interaction in the water column; interactions with sediments, suspended sediments, and colloids; and transfer to local communities of animals and plants and ultimately their toxicity. Radionuclides in seawater may behave as nonconservative substances (e.g., Pu and Am) or conservative substances (e.g., Cs, Sr, and Tc), which affects their loss to sediment and cycling in the food web ([► Radioactivity in the Marine Environment](#)).

Marine databases of radionuclide concentrations in seawater have been compiled by regional and global projects to assess the movement of ocean currents. Movement of conservative (^3H , ^{14}C , ^{90}Sr , ^{129}I , ^{137}Cs , etc.) and nonconservative radionuclides (^{238}Pu , ^{239}Pu , ^{240}Pu , ^{241}Am , etc.) have been used to study the circulation of ocean water in the Atlantic, Pacific, and Indian Oceans and interocean exchange of water. Models have been developed to describe currents such as the Great Ocean Conveyor Belt and Global Ocean Circulation Model. Radionuclide releases from Sellafield, UK and La Hague, France, have provided considerable information on the movement of water in the eastern North Atlantic and adjacent seas ([► Radionuclides as Tracers of Ocean Currents](#)).

Nuclear accidents such as the Chernobyl accident released large pulses of radionuclides (^{90}Sr and ^{137}Cs) that allow researchers to follow a pulse through the environment improving our understanding of the behavior of radionuclides and the predictability of models. The complex processes of radionuclide migration through a drainage basin are dependent on the characteristics of the specific radionuclide and the catchment. Movement of radionuclides is primarily governed by the hydrology and interactions with soil and, in aquatic systems, suspended and bottom sediment. The complex hydrological processes that control runoff such as water flow, sediment erosion, and physical and chemical processes of contaminant interactions with soil govern transport and can be described by radionuclide transfer functions per unit of deposition ([► Radionuclide Migration from Catchments, Modeling](#)). Models have been developed to predict inflow fluxes and outflow fluxes for rivers, lakes, and coastal regions. Dynamic models not only consider

migration of contaminants in water bodies, removal to deeper waters, transfer to sediment, resuspension, and transfer through the food web, but incorporate biodilution and chemical dilution of radionuclides to fish and the biotic processes that control bioaccumulation and persistence in fish ([► Nuclear Accidents, Chernobyl Fallout in Scandinavian Watersheds](#)).

Chemical speciation affects the mobility, bioavailability, and toxicity of metals and radionuclides. This is discussed for the terrestrial ([► Radionuclide Fate and Transport in Terrestrial Environments](#)), and aquatic environment ([► Speciation of Metals, Effects on Aquatic Biota](#)) and uranium ([► Uranium in the Environment: Behavior and Toxicity](#)). The marine speciation of radionuclides differs from that in freshwater ([► Radioactivity in the Marine Environment](#)). An understanding of metal/radionuclide speciation allows the prediction of the interaction of dissolved metals/radionuclides with biological surfaces of organisms. The Biotic Ligand Model combines chemistry, physiology, and toxicology simultaneously to account for geochemical speciation and binding to specific sites to assess the bioaccumulation and toxicity of the chemical species. The Biotic Ligand Model is receiving considerable attention in the study of metal toxicity to aquatic organisms, but is equally relevant to the accumulation and toxicity of radionuclides by biota ([► Speciation of Metals, Effects on Aquatic Biota](#)).

The impact of radiation on organisms depends on the exposure rate, duration of exposure, time frame of biological repair and endpoint studied, which change with age. Exposures are either acute or chronic. Acute exposures in the environment are essentially restricted to the brief period following a major accident. The major consequence of chronic exposures at low dose rates as a result of routine releases is genetic effects. Chromosomal effects are also seen in organisms including humans living in areas of naturally elevated background levels of radiation. Radiation damage to DNA may include single- or double-strand breaks. Errors in DNA repair show up as chromosomal aberrations and may eventually result in genomic instability through the generation of dicentric or acentric chromosomal fragments and may be expressed at the population level several generations later. High exposure rates affect the rates of morbidity, reproduction,

and mortality depending upon such factors as tissue sensitivity, dose fractionation, relative biological effectiveness of the radiation type, and adaptive response. Since most exposures are to low dose rates, potential effects are small and difficult to quantify and are temporary in nature, the effect being reversible. The primitive germ cell is the most sensitive cell type. The biological effect is dependent on the number of cells struck by ionizing radiation, and energy type; alpha, beta, or gamma. The life history of the organism needs to be considered in the assessment of potential effects. Most vulnerable are long-lived, slow-reproducing species. Effects generally reported in exposed populations are life shortening, reproductive impairment, increased parasite loads, decreased survival of young, increased mutation rates, deformities, and genetic effects. However, it is difficult to differentiate the effect of radiation from that of other contaminants. Protection can be at the individual, population, or community level ([► Ionizing Radiation on Nonhuman Biota, Effects of Low Levels of](#)).

Environmental Toxicology, Introduction

EDWARD A. LAWS

Louisiana State University, School of the Coast and Environment, Baton Rouge, LA, USA

Ecological Risk Assessment

Carroquino et al. ([► Environmental Toxicology: Children at Risk](#)) discuss the numerous factors that make children at greater risk from exposure to toxic substances than adults. They point out that children have higher exposures relative to body weight than adults because they drink more water (seven times more per kilogram), eat more food (three to four times as much per kilogram for children between the ages of 1 and 5), and breathe more air (up to twice as much per kilogram for a child less than 1 year old). Furthermore, during the time that their central nervous systems are developing, children are susceptible to permanent neurological damage from exposure to neurotoxins such as methyl mercury, lead, and ionizing radiation. And if

endocrine disruptors send false signals to developing reproductive organs, there is a high probability that the resulting dysfunction will be permanent and irreversible. Babies in utero are especially sensitive to the use of addictive substances such as drugs and alcohol by their mothers, and studies have shown that maternal smoking during pregnancy increases the risk of pregnancy loss, stillbirth, and infant mortality.

Willett and Foran ([► Ecological and Health Risks at Low Doses](#)) discuss what is known and, more frequently, what is not known about the mechanisms associated with human health effects caused by exposure to low doses of xenobiotics. Traditional sigmoidal dose-response models lead to the conclusion that below a certain threshold dose, there are no adverse health effects. Application of such models has led to the use of so-called no observed adverse effect levels (NOAEL), which in turn have become the basis for setting acceptable daily intakes of substances known to cause adverse effects at higher doses.

In most cases, health effects at low doses are extrapolated from experimental results at much higher doses using a linear dose-response model. But in a number of cases, low-dose treatments with a toxicant have been shown to induce a beneficial response, a phenomenon called hormesis. And pretreatment of animals with metals, specifically cadmium, copper, mercury and zinc, is protective for subsequent exposures. The mechanism underlying the acclimation to metals is synthesis of thionein, a protein that normally only occurs in trace amounts in certain tissues (blood, gills, liver, kidney, and intestine) and can effectively sequester toxic metals if the dose is not too great. But in humans, at least, there is a growing consensus that any amount of lead in the body can be damaging, and especially so in children, in whom adverse effects of lead exposure are typically associated with brain damage. Exposure to neurotoxic organophosphorus pesticides such as chlorpyrifos and to endocrine disruptors such as Bisphenol A are additional examples of risks that are poorly quantified because of inadequate understanding of the mechanisms underlying effects at low doses.

Bain ([► Ecological Risk Assessment and Animal Models](#)) discusses the many issues associated with assessing the risk to plants and animals caused by damage/modification of the environment by human activities. As noted by Bain, this is a multistep process,

culminating ultimately in a phase characterized as risk management. The first few steps involve problem formulation and analysis, and these are the focus of the chapter. Much of the information upon which current environmental standards are based has come from either acute (short-term) or chronic (long-term) exposure of organisms to stress and observing the effects on survival, growth and development, and reproduction. Since the number of organisms potentially impacted by toxic substances and stresses is very large, care must be taken to carry out these bioassays with organisms that are in some sense representative of those found in the natural environment, including in particular some of the more sensitive species and life stages. Since such bioassays in most cases involve exposure to only one stress, an alternative, particularly when concern involves a combination of stresses and toxic substances, is to expose test organisms to, for example, contaminated soil or water and observe the impacts on survival, growth, and/or reproduction. In this way, the interaction of a combination of stresses is taken into account.

Alternatives to such bioassays include use of so-called bioaccumulation factors (BAFs) to estimate the effective degree of exposure based on the concentration(s) in the organism. In the last several decades, an approach to quantifying the overall health of a community of organisms has been the use of so-called biotic indices and rapid bioassessment protocols. Such surveys compare populations and community compositions of macroinvertebrates, fish, or periphyton (communities of organism and organic matter attached to surfaces) between a reference site and the site of interest. Finally, and most recently, molecular methods, so-called “omics” (genomics, proteomics, transcriptomics, and metabolomics), have begun to be used, with the expectation that their use will provide a more sensitive assay for stress than the proverbial canary in the coal mine. While this expectation is very likely true, the interpretation of results from omics studies with respect to issues such as risk management is an evolving art.

Medema (► [Microbial Risk Assessment of Pathogens in Water](#)) discusses the theoretical and practical issues associated with using quantitative microbial risk assessment (QMRA) to estimate the risk to public health from the presence of pathogens in drinking water derived from surface sources. This is an approach

that has become popular in recent years because of the absence of relevant epidemiological data that might otherwise be used to estimate risks associated with drinking contaminated water. The use of QMRA has benefitted from the existence of a database that can be used to relate dose to the risk of infection by the protozoan pathogen *Cryptosporidium parvum*, a database, incidentally, derived from studies with human volunteers. An important point noted by Medema is that in many cases, the risk estimates are dominated by relatively infrequent so-called hazardous events, such as floods, when the water distribution system is compromised and/or the source water is seriously polluted. The water distribution system in New Orleans following Hurricane Katrina is a case in point.

The strategy in QMRA is to use mathematical and statistical models to follow drinking water from its source to a treatment system, through a distribution system, and finally to the consumer and to estimate the probability that a person drinking water from his/her tap would be infected by a particular pathogen. The process is an iterative one, and Medema notes that one of the important products of the modeling exercise is identification of knowledge gaps.

In addition to its role in estimating risks due to consumption of contaminated water, QMRA has been used to identify the risks associated with the use of reclaimed wastewater, for example, to irrigate golf courses and crops destined for human consumption, and for recreational water use. As in the case of drinking water, the strategy is to develop a probabilistic model of the public health risk associated with such practices. Persons and agencies responsible for risk management use the output of QMRA to make decisions about appropriate levels of drinking water treatment and to establish appropriate policies for water reuse and recycling.

Monitoring

Soelberg and Furlong (► [Biosensors and Bioassays for Ecological Risk Monitoring and Assessment](#)) describe a portable device for monitoring pollutants based on surface plasmon resonance (SPR) technology. This is a rather esoteric technology but simplistically takes advantage of the change in refractive index (RI) of

a solution adjacent to a gold surface when the solution contains a target compound. The gold surface is coated with so-called recognition elements that provide recognition element (e.g., antibody) attachment sites for the target compound. The technology is designed to detect targets dissolved or suspended in a liquid medium, but the chapter includes a discussion of methodologies for collecting/transferring substances/compounds from air or soil samples or from the surface of solid objects.

Details of the detection process depend on the size of the target, which may range from small molecules to single-celled organisms such as viruses and bacteria. Binding of small molecules to the sensor surface usually does not produce a change in RI large enough to be of practical use, but the detection of such small molecules can be achieved using a competition/inhibition assay involving another larger analyte. For larger molecules such as proteins, binding of the target may effect a change in RI sufficient for detection, but the signal can be amplified by addition of a second recognition element that binds to another site on the target molecule. Whole cells, which are generally too large to be directly detected by SPR technology, may, nevertheless, be detected indirectly as a result of their binding to an appropriate antibody, whose concentration is assayed by SPR in the presence and absence of the cells of interest. Because of the small size of the sensor chips, it is quite possible to design a flow-through system that can detect multiple analytes in sequential fashion. Future directions may involve the further miniaturization of the technology via illumination of an array of recognition element spots in a way that would allow detection of hundreds or perhaps thousands of analytes simultaneously.

Keum et al. (► [Biomarkers and Metabolomics, Evidence of Stress](#)) review the information and understanding of the effects of stress on organisms that has been derived from a study of metabolomics. This is an emerging area of omics research that lies downstream of genomics and proteomics. The rationale for identifying metabolic fingerprints associated with particular kinds of stress is to provide an early warning mechanism that could be used to trigger corrective action before serious damage has been done. To date, much of the metabolic work has involved single-celled microorganisms and plants and their responses to stresses

such as nutrient deprivation, lack of water, pesticides, and/or salt stress. One of the more interesting metabolomic studies has been the investigation of the response of plants to attack by herbivorous insects and pathogens. Evolutionary arguments would lead one to expect that these responses provide some degree of defense for the plants, and elucidation of the relevant upstream portion(s) of the genome has enabled genetic engineers to produce crops resistant to attack by important plant pests. As noted by the authors, the metabolomes of mammals are more complex than those of plants and microorganisms, but ultimately, lessons learned from the study of relatively simple organisms will likely facilitate more informed investigations of the human metabolome, the etiology of various forms of cancer, and the metabolic pathways of preclinical drugs being obvious topics of interest.

Boehm and Soller (► [Recreational Water Risk: Pathogens and Fecal Indicators](#)) discuss the problems associated with determining the risk of infection associated with recreational water use. Although there is a long list of pathogens that could potentially infect swimmers, more than 95% of all non-food-borne illnesses in the United States are caused by only eight: the viruses norovirus, rotavirus, and adenovirus; the bacteria *Campylobacter*, *Salmonella*, and pathogenic *E. coli*; and the protozoans *Cryptosporidium* spp. and *Giardia lamblia*. Because the concentrations of such pathogens in recreational waters are typically low, monitoring of recreational waters has traditionally relied on the concentrations of so-called indicator organisms, which are found in relatively high concentrations in human feces but are not themselves pathogenic. Problems associated with the use of fecal indicator bacteria include the variable ratios of indicators to pathogens and the fact that epidemiological studies have shown that there is a human health risk associated with swimming in waters where there is no apparent source of human fecal contamination. In such cases, the risks of infection are presumably associated with animal excreta, including birds, pigs, and cattle.

A complement and possibly alternative to traditional recreational water quality monitoring techniques is quantitative microbial risk assessment (QMRA), a health risk modeling approach that translates microbial exposures into infection or illness risk

estimates. The input to such models includes information on the die-off rates of pathogens in the environment, assumptions about the amount of water swallowed by recreational swimmers, and the infectivity of the various pathogens. Comparisons of the output of QMRA models with the incidence of relevant infections have provided valuable insights concerning the likely contribution of land runoff and animal feces to the risks associated with recreational water use. Historically, the pathogen that has received the most attention in QMRA simulations is the rotavirus because of its high infectivity and the availability of dose-response information. More recent work has focused on the norovirus following publication of relevant dose-response information in 2008.

Mechanisms

Wilson (► [Environmental Toxicology: Carcinogenesis](#)) provides an excellent overview of the numerous and diverse mechanisms responsible for causing cancer. He points out that nine different conditions must be satisfied before a malignancy can develop and that to satisfy each of these conditions, one or more mutations are needed to alter the intracellular signaling pathways and/or response to the tissue microenvironment. The human body in fact has numerous defenses against cancer, including inter alia, DNA repair systems with phenomenally high fidelity, detoxifying enzymes for reactive species, glutathione to quench reactive electrophiles, and, in the case of mammalian and human cells, the ability to become terminally senescent or undergo self-suicide (apoptosis) upon self-recognition of excess DNA damage or perturbation of cell growth controls. However, these defenses are not infallible, and a variety of mutations generally classified as (1) single-base mutations, (2) chromosomal aberrations, (3) insertions and deletions (indels), and (4) epigenetic mutations can eventually lead to cancer, the probability being directly proportional to the dose and duration of carcinogen exposure.

With a few exceptions (e.g., mesothelioma and exposure to asbestos), most cancers have multiple etiological causes. Therefore, the development of a particular form of cancer does not, in most cases, implicate a mechanism. On the other hand, the mechanisms associated with some forms of stress are rather well

established. Chronic alcohol consumption, for example, is associated with oral cavity, pharynx, larynx, esophageal, liver, and colorectal cancer, and breast cancer in women, and suspected to be involved in pancreatic and lung cancer. Alcohol (i.e., ethanol), however, is not itself genotoxic, but acetaldehyde, to which ethanol is converted in the liver, is mutagenic. Chronic alcohol use is associated with three carcinogenic pathways: DNA damage and mutation, inflammatory processes that promote proliferation, and vitamin deficiencies that perturb DNA epigenetic patterns. Exposure to ultraviolet radiation is associated with skin cancer, the most common cancer in the United States. Ultraviolet radiation in the wavelength range 290–320 nm causes DNA mutations leading to signature cyclobutane pyrimidine (6–4) pyrimidone photoproducts and pyrimidine dimers, which, if not repaired, result in signature C (cytosine) to T (thymine) and CC to TT mutations, respectively. Tobacco smoke, probably the most important anthropogenic carcinogen of the twentieth century, contains at least 300 known or suspected carcinogens and is believed to be an important etiological factor in many cases of oral cancers, esophageal, bladder, pancreatic, kidney, and possibly breast and cervical cancers. Many of the carcinogens in tobacco smoke alkylate or arylate DNA, with consequent production of promutagenic lesions, or produce DNA adducts that can lead to primarily G (guanine) to T (thymine) transversions, which are the most common single-base mutations in lung tumors. Despite the remarkable number of carcinogens in tobacco smoke, many years of smoking (~50 pack years in the case of lung cancer) are generally required before a tumor develops, a testimony, as noted by Wilson, “to the resilience of the human body to the bombardment of such a powerful combination of the carcinogens and toxic agents present in tobacco smoke.”

Jones (► [Environmental Toxicology: Oxidative Stress](#)) points out that understanding of the mechanisms that underlie oxidative stress has changed significantly in recent years. Thirty years ago, oxidative stress was understood to reflect a predominance of prooxidants (agents that initiate radical reactions, oxidize biologic components, or interfere with normal reductive and antioxidant functions) over antioxidants, leading to macromolecular damage. If oxidative stress were as simple as that, supplementation with

antioxidants would be expected to provide protection but in fact has been found to be associated with little or no health benefits in humans. Furthermore, the discovery of NADPH oxidases that generate O_2^- and H_2O_2 as signaling molecules suggested that oxidative stress could involve disruption of redox signaling and control mechanisms. Living cells have a variety of mechanisms that almost completely prevent the sort of radical chain reactions associated with the more traditional concept of oxidative stress. Although such chain reactions may be relevant in cases of acute exposure, non-radical oxidants such as H_2O_2 , lipid hydroperoxides, quinones, disulfides, and reactive nitrogen species are now believed to be far more important from the standpoint of chronic toxicity. The major mechanisms of non-radical oxidative stress involve primarily effects on thiols and selenols in proteins and secondarily mutagenic damage to DNA. Non-radical oxidants can interfere with reversible oxidation–reduction reactions of thiols or selenols (e.g., the thiol in cysteine, the thioether in methionine, and the selenol in selenocysteine) by causing abnormal oxidation or irreversible modification. These changes alter physiological function in receptor signaling, transcriptional regulation, cell proliferation, angiogenesis, and apoptosis. The hypothesized mechanisms presume that (1) all biological systems contain redox elements that function in cell signaling, (2) organization and coordination of the redox activity of these elements occurs through redox circuits dependent upon common control nodes, (3) the redox-sensitive elements are spatially and kinetically insulated, and (4) oxidative stress is a disruption of the function of these redox circuits caused by specific reaction with the redox-sensitive thiol/selenol elements, altered pathways of electron transfer, or interruption of the gating mechanisms controlling the flux through these pathways. Because biological systems are effectively engineered to conceal stress effects, subtoxic exposures can accumulate without overt signs of adverse effects. A redox proteomics/gene expression approach may therefore provide a means for systems analysis to detect evidence of oxidative stress before overt symptoms become apparent.

Laws (► [Toxic Chemical Risks](#)) note that most exposure to toxic chemicals is associated with lifestyle decisions and not, as might naively be expected,

because a person lives downwind or downstream from a source of pollution. In the United States, by far, the biggest source of exposure is tobacco smoke, which, as noted by Wilson (*vide supra*), contains several hundred known or suspected human carcinogens. The list of toxic substances includes ammonia, arsenic, benzene, benzo[a]pyrene, cadmium, carbon monoxide, nicotine, and several compounds classified as tobacco-specific N-nitrosamines. The death toll from cigarette smoking accounts for about 20% of all deaths in the United States. Fortunately, the percentage of adults who smoke in the United States has been steadily declining from 42% in 1965 to 21% in 2008–2009.

After tobacco smoke, the most significant source of exposure to a toxic chemical in the United States is alcohol (ethanol) consumption. Health problems associated with chronic alcohol consumption include liver damage, a variety of reproductive system disorders, and damage to the digestive tract. Alcohol abuse accounts for about 100,000 deaths per year in the United States. The specific causes of death include drunken driving, cirrhosis of the liver, cancer, and stroke. In addition to these direct effects on human health, alcohol abuse is associated with half of all homicides and 40% of assaults. Alcohol consumption can be addictive, and about 8% of Americans abuse alcohol or are alcoholic.

Drug misuse/abuse follows alcohol in terms of the frequency with which people are exposed in the United States. The number of deaths associated with unintentional drug overdoses has been rising more or less exponentially during the last 30 years and is now approaching 30,000 per year. The biggest single source of exposure is abuse of prescription analgesics (painkillers). The painkillers in question are opioids, which puts them in the same drug category with drugs such as heroin and cocaine. They effectively block feelings of pain by binding to opioid receptors in the brain, spinal cord, and gastrointestinal tract, but they also produce feelings of euphoria, which accounts for their abuse and compulsive use. When used as prescribed, opioid analgesics seldom result in clinical addiction. Lethal effects are associated with the fact that unprescribed opioid use can lead to a suppression of breathing that may ultimately lead to respiratory failure.

Cocaine and heroin are both illegal opioid drugs. Cocaine has virtually no beneficial application, and there is no approved medication to treat cocaine

addiction. Heroin abuse is associated with a variety of health effects related to the tendency of addicts to inject the drug with a needle that in some cases is far from sterile. Thirty-six percent of AIDS cases and an estimated 70–80% of the new hepatitis C cases in the United States each year are attributable to injection drug use. Fortunately, there are a variety of effective treatments available for heroin addiction.

Obesity has become a major public health problem in the United States and reflects an old adage about toxic substances: the dose makes the poison. Food is essential for living organisms, but consumption of too much food can lead to serious health problems, including, in the case of humans, morbidity from hypertension, abnormal amounts of fat in the blood, type II diabetes, coronary heart disease, stroke, gallbladder disease, osteoarthritis, sleep apnea, respiratory problems, endometrial, and social stigmatization and discrimination. At the present time, roughly 100 million adults in the United States are overweight or obese. The solution is to consume fewer calories, but for many people, this is a difficult goal to achieve on a sustainable basis. Behavior therapy may be necessary to effect the necessary changes in lifestyle.

Of the problems with little or no lifestyle connection, lead intoxication appears to be the most pervasive. Several decades ago, the use of leaded gasoline was a significant source of exposure, but this practice was fortunately phased out in the United States in the 1970s. Today, the major lead-related issue is the fact that many residences were painted with lead-based paint prior to 1978, when the use of such paint was banned by the Environmental Protection Agency. Unfortunately, many people live in residences that were painted prior to 1978, and at the present time, it is estimated that 250,000 children in the U.S. have blood lead levels exceeding 100 $\mu\text{g/L}$, a concentration that the Centers for Disease Control feel should not be exceeded. Effects of lead intoxication are variable and depend on the degree of exposure. In children, the biggest concern is effects on the brain. The apparently obvious solution is to remove the lead-based paint in old homes, but doing so in a safe way requires special training and appropriate equipment.

Many of the other incidents of exposure to toxic chemicals involve accidents in the home. Among the most frequent involve ill-advised mixing of cleaning

compounds, the most common examples being bleach mixed with either ammonia, vinegar, or toilet bowl cleaners. Chlorine gas, a highly toxic and corrosive gas, is often one of the products of such mixtures, and in some cases, the chemical reactions are very exothermic and can easily result in explosions.

Fate and Transport

Macaula and Hiscox ([► Airborne Toxic Chemicals](#)) use three case studies to illustrate issues related to the fate and transport of air toxics (atmospheric pollutants). Perchloroethylene (PERC) is the most common chemical associated with dry cleaning. It has been linked to liver and kidney tumors in rats and is considered to be a carcinogen by the EPA, which has ruled that the use of PERC in dry cleaning must be completely phased out by 2020. Persons living or working in close proximity to dry cleaning establishments are at greatest risk from exposure to PERC. In 2005, the EPA ruled that large dry cleaning establishments use state-of-the-art recovery systems to ensure that PERC is not emitted to the air, and that smaller dry cleaners use generally available control technologies to control emissions. Dry cleaners that operate in apartment buildings are particularly troublesome because residents are obviously living in close proximity to a source of PERC. For such establishments, the EPA has mandated the replacement of so-called transfer machines that require PERC-soaked clothes to be moved from washing to drying machines. These must be replaced by machines that both wash and dry.

Gasoline combustion creates air pollution problems on a broader scale geographically. The most troublesome issues relate to incomplete combustion of gasoline, which can result in the release of numerous pollutants. The pollutant of greatest concern from a public health standpoint is benzene, which has been linked to leukemia. Several approaches have been taken to mitigating the problems associated with incomplete gasoline combustion. First, many states now require gasoline pumps to be equipped with vapor recovery nozzles in order to trap gases that evaporate and would otherwise escape during refueling. Second, gasoline has been reformulated, primarily by addition of ethanol, to burn more completely. This reformulation has significantly reduced carbon monoxide emissions, a cause of

smog, but at the same time has led, in the case of California, to a 54% reduction in benzene emissions. A longer-term solution may be electric cars, which could be recharged at night when electric power demand is relatively low. However, such a transformative change would require an unprecedented degree of cooperation between the electric power and automobile industries.

Mercury emissions are an even longer-range problem, with emissions in Asia linked to mercury levels in North America. Although historically, industrial use of mercury has led to some very serious localized human health problems (e.g., Minamata Disease), on a global scale, coal combustion accounts for about 60% of atmospheric emissions. Inhalation of mercury vapor can lead to serious health effects because the mercury reaches the brain, where it can cause irreparable neurological damage. Ironically, the primary post-combustion technologies that reduce mercury emissions do not specifically target mercury but instead are designed to remove oxides of sulfur and nitrogen. In the USA, at least these stack gas-scrubbing technologies reduce mercury emissions by about 33%. In recent years, the USA has considered controlling the remaining mercury emissions from power plants using a cap-and-trade approach, but this idea was struck down by a Court of Appeals ruling. It now appears that mercury emissions from power plants will be regulated by the requirement for use of Maximum Achievable Control Technology.

Oliver and Heathwaite (Invisible Threats: Transfer of Pathogens and Nutrients Through and Across Agricultural Soils) discuss the tendency of agricultural soils to serve as conduits for pathogens and nutrients derived from agricultural activities. The nutrients of primary concern are nitrogen and phosphorus. The pathogens include a somewhat longer list of organisms (*E. coli* O157, *Salmonella* spp., *Campylobacter jejuni*, *Listeria monocytogenes*, *Cryptosporidium parvum*, and *Giardia intestinalis*) that are shed by livestock and are, coincidentally, pathogenic to humans. The ability of soils to sequester N is very much influenced by the fact that soil particles typically have a negative surface charge. Thus, the ammonium ion, NH_4^+ , is effectively retained in soils, while nitrate, NO_3^- , is not. Phosphate, although negatively charged, is typically retained in soils because of its tendency to bind with ferric iron

(Fe^{3+}), aluminum (Al^{3+}), or calcium (Ca^{2+}). The survival and movement of pathogens is very much influenced by the water content of the soil and the existence of conduits (e.g., macropores) for transport within the soil matrix. Under appropriate conditions, pathogens may survive within soils for literally months before being flushed out by seepage from storm events. Strategies such as no-till agriculture that are intended to minimize the use of pesticides and reduce nutrient runoff may actually exacerbate pathogen export by maintaining the continuity of macropores. In some cases, relatively simple strategies can be used to minimize pathogen problems, a case in point being the application of manure in a broadcast slurry to effect a more rapid destruction of associated bacteria through UV radiation and desiccation. Because of the potentially long time that pathogens and particularly nutrients may be sequestered in soils before being released, a considerable lag in time may exist between the implementation of management practices and discernible changes in the water quality of drainage systems. Awareness of this time lag is important to the informed assessment of management practices.

Bienfang et al. ([▶ Bioaccumulation/Biomagnifications in Food Chains](#)) discuss two somewhat different case studies that illustrate problems associated with the bioaccumulation and/or biomagnifications of toxic substances in food chains. Ciguatera fish poisoning (CFP) is the most common food-borne disease related to the consumption of marine finfish, with ~50,000 reported cases each year, but the actual number may be much larger. The toxin responsible for causing CFP comes in several dozen congeners. The various forms of ciguatoxin (CTX) are believed to be metabolites of gambiertoxins produced by various species of the dinoflagellate genus *Gambierdiscus*. Because CTX is toxic at very low concentrations, and in part because there are so many congeners, analytical detection of CTX has been problematic. In fact, virtually all reported cases of CFP are based on symptomology rather than detection of CTX. Short-term effects of CFP are a variety of gastrointestinal disturbances. The long-term and more troublesome effects are of a neurological nature.

As a high-molecular-weight lipid, CTX is a prime candidate for biomagnifications, but because of the aforementioned analytical detection issues, direct evidence of biomagnification of CTX is lacking.

The indirect evidence is the fact that the greatest number of reported cases of CFP and the most severe reported symptoms typically involve consumption of carnivorous as opposed to herbivorous fish.

Mercury is a well-known neurotoxin that is found in the environment in several different forms, by far the most troublesome of which is methyl mercury. In contrast to other forms of mercury that are rather efficiently eliminated from the human body, methyl mercury is, at least initially, retained with about 95% efficiency following consumption and is only slowly excreted at a rate of ~1% per day. The damaging effects of methyl mercury are associated with its ability to cross the blood-brain and placental barriers. As a result, exposure to methyl mercury can cause serious and permanent neurological damage. Fetuses appear to be particularly susceptible, as women evidencing no overt symptoms of mercury intoxication have been known to give birth to babies hopelessly brain damaged from the effects of methyl mercury.

Though methyl mercury has been reported at high concentrations in some fish populations, the documented cases of methyl mercury intoxication have all involved anthropogenic discharges. Natural concentrations of methyl mercury in fish are, with very few exceptions, invariably accompanied by even higher concentrations of selenium, the only exceptions being pilot whales and mako sharks. Evidence that has accumulated during the last 50 years has shown that mercury and selenium effectively sequester one another. The absence of any apparent mercury-related adverse health effects associated with the consumption of virtually any species of marine fish reflects the fact that any mercury in the fish is rendered inert by the selenium in the same fish.

Prevention

Kennedy and Tierney ([► Xenobiotic Protection/Resistance Mechanisms in Organisms](#)) point out that the toxic effects of xenobiotics can be mitigated either through acclimation or adaptation, the former involving intragenerational tolerance and the latter involving intergenerational resistance. Acclimation may involve nothing more complicated than avoidance, in some cases instinctive and in other cases the result of experiential learning. Adaptation involves selection for

individuals in a population that for some reason are less sensitive/more resistant to a xenobiotic than other members of the population. Generally speaking, adaptation involves either (1) toxicokinetically derived mechanisms, which alter the way in which organisms absorb, biotransform, and excrete chemicals, and/or (2) toxicodynamically derived mechanisms, in which target sites are modified to reduce sensitivity. An important consideration in the case of toxicokinetic mechanisms is the fact that lipophilic chemicals such as PCBs and chlorinated hydrocarbons are not easily excreted because they partition back into cells and tissues from excretory media (e.g., urine, bile). Water-soluble chemicals are not reabsorbed because lipid membranes of cells lining excretory routes act as barriers to their reuptake. Thus, in the case of lipophilic toxins, toxicokinetic modes of adaptation involve transformation of the toxin into a more polar, water-soluble compound. This is generally accomplished through a series of reactions characterized as Phase I and Phase II. In Phase I reactions, a nucleophilic functional group is introduced or exposed in the parent molecule, and in Phase II, the Phase I metabolite or parent molecule already containing a functional group is conjugated to an endogenous molecule to form a water-soluble product. Cytochrome P-450 mixed-function oxidases, for example, are a superfamily of heme enzymes found in all living species that mediate Phase I transformations. Phase II involves conjugation reactions that produce substrates that are secreted with great efficiency. Glutathione conjugation mediated by glutathione S-transferases is a good example of a Phase II reaction.

Toxicodynamically derived resistance refers to alterations in xenobiotic-receptor interactions, which can be effected by alterations in target site, increased or decreased concentrations of target molecules, or circumvention of target function. Structural changes in the target site can significantly reduce binding of the xenobiotic but, as noted by the authors, can lead to biological effects from reduced functionality of the target in its normal physiological roles. So-called knockdown resistance, i.e., insensitivity of Na⁺ channels to insecticide (e.g., DDT and pyrethroids) inhibition, may occur by one or another of several mechanisms, including a reduction in the number of Na⁺ channels, changes in the fluidity of nerve membranes, or alterations in the binding characteristics

between Na^+ channels and insecticides. In the case of citrus-scale mites, an interesting adaptation to HCN, which interferes with cytochrome oxidase in the electron transport chain, is the use of an HCN-insensitive flavoprotein as an alternate electron carrier in the place of cytochrome oxidase. Application of molecular tools such as toxicogenomics, bioinformatics, and metabolomics has the potential to greatly enhance understanding of the mechanisms by which organisms, including humans, acclimate and adapt to xenobiotics.

Consumption of fish and shellfish is associated with a number of well-documented beneficial health effects, but Meujo and Hamann (► [Science, Policy, and Risk Management: Case of Seafood Safety](#)) point out that there are also health risks associated with fish and shellfish consumption. The risks reflect the fact that fish and shellfish may contain dangerous concentrations of pathogenic bacteria and viruses, parasites, toxic metals, pesticides, drugs, and a variety of toxins. In the United States, the Food and Drug Administration (FDA) is primarily responsible for ensuring that seafood is safe to eat, and the strategies taken by the FDA to fulfill this mission reflect the diverse sources of the contaminants of concern, practical issues associated with analytical detection and surveillance, and the extent to which postharvest processing can be used to reduce contaminant concentrations to an acceptable level.

Bacterial and viral pathogens can be very effectively eliminated by a process called high hydrostatic pressure (HHP) treatment. However, neither fish nor shellfish survive HHP, and that fact is a deal-killer for oysters destined for the raw oyster market. Two promising alternatives to HHP are X-ray treatment and supercritical CO_2 exposure. Both methods appear capable of killing the relevant pathogens (e.g., *Vibrio parahaemolyticus*) without killing the oysters.

Neurotoxins produced by a variety of microalgae can accumulate in fish and shellfish to levels that cause serious health effects on consumers. The most troublesome of these natural neurotoxins is ciguatoxin, but there is a long list of others, including brevetoxin, okadaic acid, and domoic acid. An additional neurotoxin, scombrototoxin, is not produced by microalgae but instead is a product of a catalytic reaction involving the conversion of histidine into histamine. One of the problematic issues with these neurotoxins is the fact that they

are heat stable, so cooking has little or no effect on the toxicity of seafood containing these compounds. From the standpoint of prevention, what is badly needed in the case of these neurotoxins is a simple and inexpensive test to determine whether seafood is safe to eat. In the case of brevetoxin, an antibody–antigen assay using ELISA (enzyme-linked immunosorbent assay) has recently been reported. The development of similar high-specificity probes for the other troublesome neurotoxins would clearly be desirable.

Correctives

Portier (► [Bioremediation and Mitigation](#)) describes several case studies in which heterotrophic microorganisms have been used to clean up organic waste in soil and water. The metabolic pathways involved in the biological degradation of various classes of organic compounds are reasonably well understood, and the ease/speed with which microbes can decompose different kinds of organic wastes have been the subject of numerous investigations. Normal alkanes, for example, are oxidized to primary alcohols, which are then further oxidized to carboxylic acids. The carboxylic acids are then converted via beta-oxidation enzymes to acetic acid and a simpler alcohol. Polycyclic aromatic hydrocarbons (PAHs) are generally considered to be the most toxic components of petroleum. Simple PAHs are initially converted via dioxygenase enzymes to catechols and substituted catechols. The strategy in bioremediation is to identify a community of microorganisms with metabolic pathways that enable them to effect such transformations and, insofar as possible, to create environmental conditions that encourage their catabolism of toxic substrates.

The examples cited by Portier illustrate some of the diverse strategies that may be employed to effect bioremediation. The first example concerns the application of adapted microbial consortia to remediate diesel-contaminated soils at an active rail yard. Total petroleum hydrocarbon concentrations in the soil were reduced by almost a factor of 20 over a 6-month timeframe. In the second example, an immobilized packed bed reactor filled with support media was used to provide a habitat for adapted microbes with enzyme systems capable of degrading wood preservative contaminants in groundwater. The contaminated

groundwater was pumped through the reactors and then used to irrigate surface vegetation and/or percolate back into the ground. The site in question is now zoned recreational, and groundwater is close to meeting federal maximum contaminant level water quality criteria. The third example involves in situ remediation of groundwater via use of soil seeding bioreactors called bioplugs or bioconduits. Through manipulation of subsurface hydrology, the contaminated groundwater is routed through the bioreactors. Surplus microbial biomass is allowed to escape into the surrounding soil, where the microbes infiltrate contaminated areas and further reduce the concentrations of toxic chemical constituents. Although bioremediation is not an option for all sorts of pollutant (e.g., heavy metals), the strategy has proven to be remarkably successful in cleaning up biodegradable wastes in terms of both cost and sustainability of treatment effects.

Jewett and Wascom (► [CERCLA, Sustainability and Public and Environmental Health](#)) discuss another kind of corrective strategy: legal action. The particular legal subject is the Comprehensive Environmental Response, Compensation, and Liability Act (CERCLA), one of the last pieces of legislation signed into effect during the administration of President Jimmy Carter. CERCLA was intended to provide a mechanism to clean up “facilities” contaminated by inappropriate disposal of hazardous substances, with the word facility very broadly defined. Hazardous substances are also broadly defined, with one notable exception: petroleum and petroleum products, natural gas, and synthetic gas used for fuel. This is the so-called Petroleum Exception.

Cleanup activities carried out under the “Superfund” component of CERCLA were originally financed by taxes imposed on the chemical and petroleum industries and an environmental tax on corporations. More recently, Superfund has been supported by allocations from the federal general fund and through additional legislation, such as the American Recovery and Reinvestment Act.

Thirty years after its passage, the report card on CERCLA is mixed. There have been some very noteworthy success stories. Of the more than 1,600 sites at one time or another on the National Priority List (NPL), 340 have been delisted, i.e., cleaned up. The case study discussed by Jewett and Wascom, the Agricultural Street Landfill in New Orleans, is a provocative

example given the apparently sweeping powers the EPA has to investigate and correct problems at facilities contaminated by inappropriate disposal of hazardous substances. One certainly has to ask why a municipality would elect to locate residential housing and an elementary school on top of a landfill known, during its operative days, as Dante’s Inferno. A hazard ranking score (HRS) of 28.5 is sufficient to place a facility on the NPL. The HRS for the Agricultural Street Landfill was 50.0. Why were the results of tests conducted at the site in the 1970s never made public? And why, when the presence of lead, mercury, and arsenic was detected in the soil, was a clay barrier not constructed (as recommended) to contain the contamination prior to the construction of an elementary school?

These are troublesome questions, but there is another side to the coin. The Small Business Liability Relief and Brownfields Revitalization Act has provided a defense for parties who have generated, transported, or arranged for transport of only small amounts of hazardous material and for so-called innocent landowners in the case of facilities acquired through inheritance or bequest. Such revisions have reshaped CERCLA to be more effective and fair, but, as noted by Jewett and Wascom, by no means supplant CERCLA. In summary, CERCLA is a piece of legislation with a very ambitious goal, and as time goes by may well be modified to make it more effective and fair. The need for such legislation is clearly apparent now and is unlikely to go away in the future.

Environmental Toxicology: Carcinogenesis

VINCENT L. WILSON

Department of Environmental Sciences, Louisiana State University, Baton Rouge, LA, USA

Article Outline

- Glossary
- Definition of Subject
- Introduction
- Accumulation of Mutations
- Properties of Cancer Cells

Oncogenic Mutations
 Cytotoxicity and Cellular Turnover
 Genomic Instability
 Susceptible Individuals
 Dose and Complex Dose
 Environmental Carcinogens
 Future Directions
 Bibliography

Glossary

Complex dose The biologically effective (e.g., carcinogenic) dose of an agent that takes into account additive, subtractive, and synergistic interactions in a complex mixture

DNA adduct Any covalent addition or modification of a nucleotide or the phosphate backbone in DNA

Driver mutation Oncogenic mutations that are primary to the further advancement and development of a cancer

Genotoxic Causes DNA damage and/or mutations

Immortal stranding The retention of a specific strand of the DNA double helix by a stem cell regardless of how many times the cell divides

Initiation The formation of the first oncogenic mutation in a cell

Oncogenic process Any process that is involved in the development of a tumor

Proto-oncogene A normal gene that upon mutation produces an active oncogenic component in the development of cancer

Promotion The effect of encouraging the development of cancer by increasing cellular proliferation and turnover

Tumor suppressor gene A normal gene whose expression product appears to block the development of cancer

Definition of the Subject

Carcinogenesis is the induction of cancer by exposure to exogenous agents, chemical or physical carcinogens. Cancer is a large family of life-threatening environmental diseases. This definition is supported by the plethora of published evidence that cancer results from the accumulation of DNA mutations, which in turn depends upon progressive cell divisions. This process takes time. The accumulation of a sufficient allotment

of oncogenic genetic errors for cancer to develop occurs over a large number of cell divisions (10s to 1,000s), which may take 10–50 years or more. Even without the intentional introduction of carcinogens into the human body by occupation or habits such as tobacco smoking, alcoholic use, and other life-style choices, our environment provides numerous opportunities and sources of DNA-damaging and potentially mutagenic lesions in our cells. Given enough time, the risk of cancer is inherent with life, while exposure(s) to carcinogenic agents not only shortens the time required to achieve tumor development, but also increases the probability of cancer occurrence.

In the present document, the term environmental carcinogenesis encompasses the induction of cancer by exposure to carcinogenic agents whether natural or man-made, present in our environment. The term environmental carcinogens will refer to all compounds foreign to the human or mammalian body, that is, xenobiotic and physical carcinogens. However, the development of cancer is a long and involved process that requires numerous cellular perturbations generally produced by both exogenous and endogenous reactive agents. As will be clarified in this chapter, for many cancers it is exceedingly difficult to separate out the contributions of the various etiological factors, including natural environmental carcinogens from man-made chemicals (see Table 1). For instance, a number of carcinogenic xenobiotics have been identified that are inherent in our food supply, for which the best-known examples are aflatoxins. Aflatoxins are present in fresh corn, peanuts (and peanut butter), and other grains, while other carcinogens such as PhIP (2-Amino-1-methyl-6-phenylimidazo[4,5-*b*]pyridine), are produced from cooking meat [1, 2]. The potent nephrotoxic aristolochic acid component of certain herbs sometimes used in herbal remedies produces urinary tract cancers [3–5]. To add to this disparaging viewpoint, the lack of proper nutrition, such as folate deficiency, also leads to genomic instabilities that are involved in carcinogenesis [6–8].

The contribution of a few environmental carcinogens is more clearly evident in selective cancer cases. Skin cancer induced by the ultraviolet (UV) radiation from sunlight is one of the more prominent environmental cancers in our world [9, 10]. Mesothelioma may result from inhalation of asbestos fibers [11].

Environmental Toxicology: Carcinogenesis. Table 1 The causes of cancer

Agent//Source	Examples	Associated cancer(s)
Chemical carcinogens ^a	Aflatoxins	Hepatocellular carcinoma
	Alcohol	Hepatic, breast, stomach, colorectal, oral, ovarian, pancreatic
	Aristolochic Acid	Urinary tract transitional cell carcinoma (renal, ureter, bladder)
	Asbestos	Mesothelioma and lung cancers
	Benzo(a)pyrene	Lung cancers (tobacco smoke associated)
	Benzene	AML and non-Hodgkin's lymphoma
	Formaldehyde	Nasopharyngeal
	Metals	
	Arsenic	Skin, liver, lung, kidney, bladder
	Beryllium	Lung cancer
	Cadmium	Lung
	Chromium (VI)	Lung, stomach
	Nickel compounds	Lung, nasopharyngeal
	Asbestos	Mesothelioma and lung cancers
Physical carcinogens ^a	Ionizing radiation (α , β , γ , X-rays, etc.)	multiple types of tumors
	Sunlight (Solar Radiation)	Skin (BCC, SCC, and melanoma)
Biological carcinogens	Hepatitis B, chronic infection	Hepatocellular carcinoma
	Hepatitis C, chronic infection	Hepatocellular carcinoma
	Epstein-barr virus	Burkitt's lymphoma, nasopharyngeal carcinoma
	HPV (high risk types)	Cervical cancer
	Schistosoma hematobium	Bladder cancer
Parasites (chronic inflammation)		
Chronic inflammation	Inflammatory bowel disease Crohn's disease Ulcerative colitis	Colon or intestinal cancers
	Chronic pancreatitis	Pancreatic cancer
	Barrett's esophagus	Esophageal carcinoma
	Helicobacter pylori, chronic gastritis	Stomach cancer
Fidelity limits of enzymatic processes	DNA Replication errors (1 to 3 errors per genome replication) ^b	Possibly all or most cancers
	Imperfect fidelity of DNA repair systems	Possibly all or most cancers

Environmental Toxicology: Carcinogenesis. Table 1 (Continued)

Agent/Source	Examples	Associated cancer(s)
Genetics (inherited predisposition)		
Cancer family syndromes	Breast cancer family (BRCA1 or BRCA2 gene mutation)	Breast cancer
	FAP (Familial adenomatous polyposis) (APC gene mutation)	Colorectal carcinoma
	Hereditary melanoma syndromes Dysplastic nevus syndrome Familial atypical mole-malignant melanoma syndrome Melanoma-astrocytoma syndrome	Melanoma
	HNPCC (Hereditary nonpolyposis colorectal carcinoma) (MSH2 or MLH1 gene mutation)	Colorectal carcinoma
	Li-Fraumeni cancer family syndrome (TP53 gene mutation in 70% cases)	Osteosarcoma, adrenocortical carcinoma, brain, breast, colon, leukemia, lung, pancreatic cancers, various sarcomas
	MEN2A and 2B (Multiple endocrine neoplasia, types 2A and 2B) (RET gene mutation)	Pheochromocytoma, thyroid carcinoma
	Neurofibromatosis types 1 and 2 (NF1 and NF2 gene mutations)	Neurofibromas, schwannomas, meningiomas, ependymomas, acoustic neuromas
	Retinoblastoma (RB gene mutation)	Retinoblastoma
	Von Hippel–Lindau Syndrome (VHL gene mutation)	Renal cell carcinoma, pheochromocytoma, hemangioblastomas
Other genetic diseases	Ataxia telangiectasia (AT)	Leukemias, lymphomas
	Bloom's syndrome	Leukemias, lymphomas, carcinomas
	Xeroderma pigmentosum (XP)	Skin (BCC, SCC, and melanoma)

^aAll chemical and physical carcinogens listed are classified as Group I: Carcinogenic to Humans, by the International Agency for Cancer Research (IARC), update January 16, 2009; <http://monographs.iarc.fr/ENG/Classification/index.php>

^bDNA synthesis and the replication machinery has an astonishing fidelity of one error per 10^9 base pairs, which results in one to three errors every time the human genome is replicated [31, 32].

Exposures due to personal life-style choices and habits, such as tobacco smoking and alcohol use have been well studied and also provide clear examples of environmental carcinogenesis [12–16]. The carcinogenic impact of occupational exposures is another area that offers substantial clarification of mechanisms and risks involved in cancer development. Similarly, medical exposures, especially those involving chemotherapeutic agents and radiation therapy provide further risks of carcinogenic impacts, although these agents comprise risks of a spectrum of cancer types.

Since the same principles and mechanisms involved in the development of human cancer apply to various animal species, and the primary focus of environmental carcinogenic risk calculations is designed to protect humans, the human will be the primary focus of this chapter on environmental carcinogenesis.

Introduction

The biophysical aspects of this planet support and sustain life, while also being conducive to the

development of cancer in most multicellular species. Fortunately, the risk of cancer is low without the assistance of acute exposures to carcinogenic agents. However, the risk and incidence of cancer increases with age in the mammalian organism [17, 18]. The adage that “cancer is a disease of old age” has its foundations in empirical observation, while also demonstrating the requirement of extensive cell divisions (10s to 1,000s) for the development of tumors [19, 20]. The mixtures of gases that make up Earth’s atmosphere sustain the carbon-based metazoan life forms on this planet due to the cellular oxidative metabolic processes. Low carbon dioxide (less than 0.04%) and high oxygen (20.9%) is crucial to the life-sustaining respiration of mammals. Unfortunately, the same life-sustaining oxidative metabolism functioning in the human cell also produces reactive oxygen species (ROS) that damage our cellular components including membranes, lipids, microfilaments, enzymes, and DNA. Additional ROS are produced by exposures to exogenous toxicants, including ionizing radiation, UV light, numerous toxic chemicals and heavy metals. The level of oxidative damage to mammalian DNA has been reported to be as high as 10,000 insults per cell per day [21, 22]. DNA replication (cell division) in the presence of oxidative damage leads to fixed mutations (permanent mutations). Endogenous DNA damage occurs in the form of modified nucleotides (e.g., 8-hydroxy-2'-deoxyguanosine, thymidine glycol, 5-hydroxymethyluracil, etc.), depurination, depyrimidination, deamination, and single and double strand breaks on a daily basis from oxidative processes [23–27]. Tallying up all these insults suggests that human DNA receives greater than 10,000 insults per cell per day. Other mammals receive similar insults at the level proportional to their basal metabolic rates, that is, mice and rats receive approximately tenfold more DNA damage per cell per day as humans. The DNA repair systems of the human cell have a phenomenally high fidelity, but even an error rate of only 0.0001% (the equivalent of one in a million) may lead to the accumulation of genetic errors on the order of 1 every 100 days for each and every cell [28, 29]. To add further insult to injury, the fidelities of DNA polymerases are also less than perfect, although remarkably precise in the human, such that errors are incorporated during DNA synthesis [30]. The reported occurrence of

one to three mutations every time the human genome is duplicated further supports the limitations of cellular functions and endurances [31, 32]. Thus the processes involved in sustaining life also lead to DNA mutations. These endogenous life processes coupled with the environment that is so crucial in supporting and nurturing life leads to the accumulation of DNA mutations and, over many years, a significantly increased risk of cancer.

Nature maintains a balance between productive life and cancer risk. As with any genetic disease that terminates life before reaching a reproductive age or during the fertile years, the early occurrence of terminal cancer would be self-limiting. Barriers to cancer development are necessary for the survival and proliferation of the species. As advanced life forms evolved, so did a number of protective mechanisms that are quite effective in derailing tumor development. In addition to detoxifying enzymes for reactive species, glutathione to quench reactive electrophiles, and numerous DNA repair mechanisms, mammalian and human cells become terminally senescent or undergo self-suicide (apoptosis) upon self-recognition of excess DNA damage or perturbation of cell growth controls [33]. The processes involved in cellular aging, that is, loss of telomere length of chromosomes, shifts in gene expression, and responsiveness to external signals that finally lead to replicative senescence after a finite number of cell divisions, may actually be a means to block the development of cancer. Terminal differentiation programs, such as the short 4–5 day lifespan of the colonic mucosal cells, and the approximately 14 day lifespan of the squamous differentiating skin keratinocytes ensures that these cells will not be around if they pick up too many oncogenic DNA mutations during the rapid cell divisions that take place immediately prior to the final terminal differentiation processes [34]. Should these intracellular anticancer mechanisms fail, the local tissue microenvironment plays a role in containing or dampening the expansion of abnormal cells, and the body’s immune system may take over and remove aberrant cells, thereby removing the threat of cancer [35–37]. Considering the low frequency of occurrence of cancer until old age, these endogenous protective anticancer mechanisms are advantageous for the human race, albeit not necessarily for the less fortunate individual statistical outlier.

Accumulation of Mutations

Avenues are available for the determined cell to find ways to circumvent these protective barriers against cancer development. With each protective mechanism, one or more DNA mutations enable cells to bypass this blockade and progress toward a neoplastic state. These “driver” mutations are the primary oncogenic errors involved in tumor development. Due to the number and variety of checks and balances opposing oncogenic development, a cell must accumulate a sufficient number of “driver” mutations to circumvent all the blockades for a tumor to develop, without meeting its demise by one of these protective mechanisms. This process is a statistical probability that weighs heavily on the healthy life side until old age, unless perturbed by carcinogenic exposures whereby the probability of tumor development increases (see [Table 1](#)). The rate of tumor development is increased sometimes many orders of magnitude above the spontaneous background rate by exposure to exogenous carcinogens. The timing of the onset of cancer is due to the dose and duration of carcinogen exposure and not to either chronological age or the age that exposure began [[38, 39](#)].

Properties of Cancer Cells

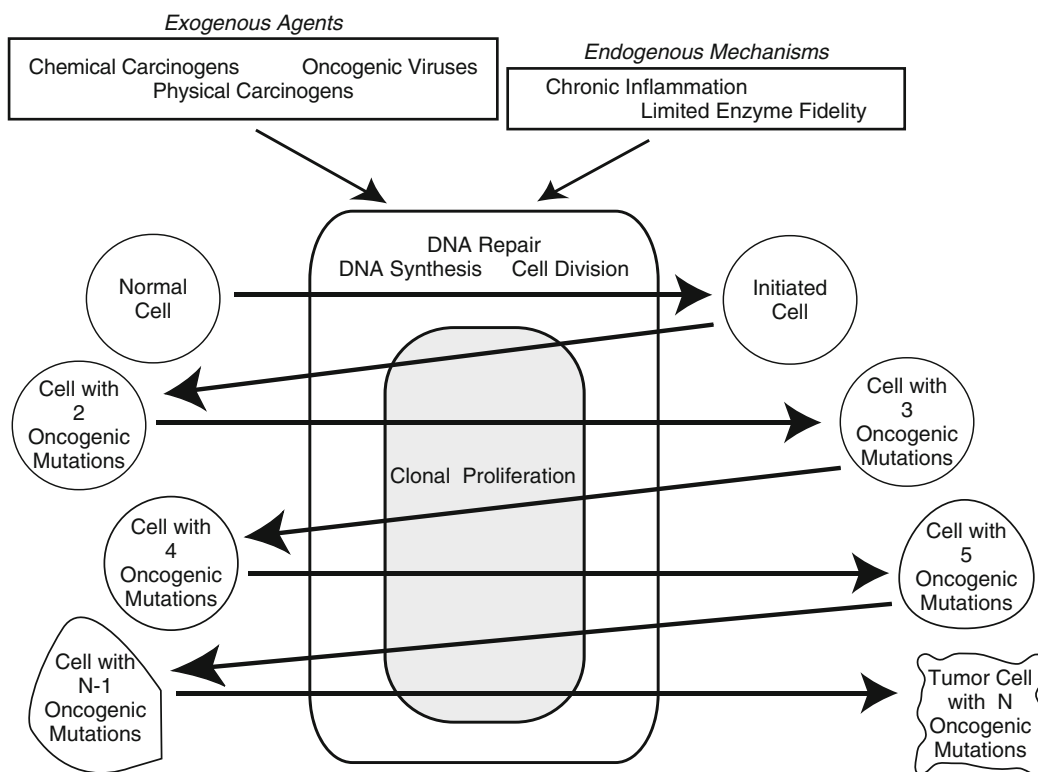
The product of carcinogenesis is a thriving cancer cell. The successful cancer cell has collected driver mutations that enable it to: (1) ignore antigrowth signals in the tissue microenvironment; (2) be at least partially independent of the requirements for exogenous mitogenic factors for cellular growth and proliferation; (3) block or inhibit differentiation programs; (4) evade apoptotic and senescence programs; (5) establish an extended replicative potential or immortality; (6) secrete factors that enhance and sustain angiogenesis; (7) acquire the ability to evade the immune system; and (8) acquire genomic instability [[33, 35–37](#)]. These are the minimal requirements for the development of a benign tumor. A malignancy has acquired one or more additional mutations that confer a ninth property, that of invasive and metastatic capabilities. For each of these nine properties of cancer, one or more mutations are needed to alter the intracellular signaling pathways and/or response to the tissue microenvironment. In some cases, only one driver mutation may be required to confer a specific property, but many of

these nine properties are conferred only after multiple oncogenic mutations have been established in the developing cancer cell genome.

Oncogenic Mutations

The establishment of each driver mutation requires the successful negotiation of several protective molecular mechanisms. First of all, the formation of a DNA adduct, single or double strand break, or other form of DNA damage must occur due to exposure to a physical carcinogen (i.e., UV light or ionizing radiation), a chemical carcinogen, or the presence of an endogenous reactive species that has successfully negotiated the gamut of detoxifying enzymes and glutathione. This DNA damage must also have occurred in an oncogenically important base, gene, and/or sequence. Next this DNA damage must be converted into a permanent genetic alteration, a fixed mutation. The formation of a fixed mutation requires the combined failure of the cell's genome surveillance system to recognize and/or repair the DNA lesion, incorporation of an incorrect nucleotide (miss-pairing), and the survival of this mistake up to and through a second round of DNA synthesis, or through faulty repair [[40–42](#)] ([Fig. 1](#)). The frequency by which this occurs is exceedingly rare, as noted above by the high fidelity of the DNA repair system alone. The bombardment of cells in the human body by acute and chronic exposures to exogenous carcinogens increases the frequency of DNA damaging events and cellular turnover, both of which in turn increase the occurrence of fixed mutations and tips the balance of these processes toward an increased risk of cancer development ([Fig. 1](#)). The higher the dose received and the longer the exposure, the greater the risk of cancer.

There are several mechanisms by which carcinogens can influence the formation of DNA mutations, and the type of mutation(s) formed is dependent upon the mode of action of the carcinogen (see [Table 2](#)). Exogenous and endogenous agents can produce DNA mutations by directly damaging DNA, perturbing DNA replication machinery, polymerases, etc., inhibiting DNA repair systems, critically shortening telomere lengths, altering the cytoskeletal proteins, mitotic spindle fibers, enzymes involved in mitotic processes, and by stimulating cellular turnover and proliferation.



Environmental Toxicology: Carcinogenesis. Figure 1

A schematic representation of the accumulation of oncogenic mutations and the influences of exogenous agents and endogenous mechanisms on this process. Each time a cell completes the cell cycle and divides, there is a good chance that one or more mutations will occur. Fortunately, most of these DNA mutations are not oncogenic. However, the probability of an initiated cell or a cell with two or more oncogenic mutations, picking up another oncogenic mutation increases as clonal expansion occurs. A tumor develops once a sufficient number of oncogenic mutations have accumulated to bypass all of the protective barriers of the normal human cell

These different mechanistic impacts on cells encourage the formation of a variety of mutations, generally classified as single base mutations, chromosomal aberrations, insertions and deletions (indels), and epigenetic mutations.

Single Base Mutations

Single base mutations encompassing single base substitutions and single base indels represent a significant proportion of cancer-associated genetic errors. It is generally accepted that the more transformed a cell becomes, the more likely a single base substitution will occur in an activating site of a proto-oncogene or

in deactivating site in a tumor suppressor gene. Most of the alkylating and arylating carcinogens listed in [Table 1](#) are known to produce single base mutations, especially single base substitution mutations. Oxidation and the production of ROS and reactive nitrogen species (RNS), common with inflammatory processes, are also well known to produce single base mutations [25, 43]. The vast majority of errors produced by DNA polymerases and the DNA replication machinery are single base mutations [31, 32].

Single base substitution (SBS) mutations represent 85% or more of the activating and deactivating mutations in proto-oncogenes and tumor suppressor genes, respectively, in human tumors [44–48]. Although

Environmental Toxicology: Carcinogenesis. Table 2 Genotoxic and carcinogenic mechanisms

Mode of action	Effects	Mutations	Instability
1. Direct DNA damage (covalent alterations)	Base pair mismatch	Single base mutation (SBS and Indels)	NSI
	Depurination and depyrimidination	Indels (larger than one base) Epigenetic mutations	EIS
	Single strand break	Chromosomal aberrations (deletions, insertions, translocations)	CIN
	Double strand break		
2. Interference with DNA polymerases and replication machinery	Base pair mismatch	Single base mutation (SBS and Indels)	NSI
	Microsatellite slippage	Microsatellite length changes	MSI
	Perturbation of epigenetic patterns	Epigenetic mutations	EIS
3. Interference with cytoskeleton and mitotic spindle Mitotic proteins and enzymes	Nondisjunction	Aneuploidy	CIN
4. Interference with DNA repair systems	Faulty mismatch repair	Microsatellite length changes	MSI
	Improperly or unrepaired	Single base mutation Indels	NSI
	Single and double strand breaks	Chromosomal aberrations (translocation, inversions, deletions, etc.)	CIN
5. Interference with maintenance of telomeres	Critically shortened telomeres	Dicentric chromosomes aneuploidy	CIN
6. Cytotoxicity	Increased cell turnover	(Most types of mutations)	Any possible (NIS, MSI, CIN, EIS)
	Increased opportunity for accumulating mutations		
7. Stimulate cellular proliferation (promotion)	Clonal expansion	(Most types of mutations)	Any possible (NIS, MSI, CIN EIS)
	Increased probability for accumulating mutations		

CIN chromosomal instability, *EIS* epigenetic instability, *MSI* microsatellite instability, *NIS* nucleotide instability (also known as single base instability)

single base deletions and insertions are sometimes involved, SBSs represent the vast majority of these oncogenic errors [47, 49, 50]. This has been clearly demonstrated in human tissues for a number of proto-oncogenes and tumor suppressor genes, for example, Braf, H-, K-, and N-Ras, RB and TP53 [45–47, 49–51]. Missense SBS mutations alone comprise more than 75% of all identified disease-associated TP53 mutations [49].

Chromosomal Aberrations

Cancer-associated chromosomal errors have been known since the middle of the last century. Cytogenetic studies identified numerous leukemia-specific (and lymphoma-specific) chromosomal translocations [52]. The best-known example is the chronic myeloblastic leukemia (CML)-specific chromosomal translocation between chromosomes 9 and 22 [t(9;22)(q34;q11)], which forms the chimeric bcr-abl gene. Chromosomal

errors are prominent in many types of human tumors and most likely crucial to the development of at least some leukemias and lymphomas [52]. Aneuploidy is the most frequently identified mutation in human cancers, albeit cytogenetic analysis is routine medical care for cancers.

Disruption of cellular cytoskeleton, mitotic spindle apparatus, and/or enzymes involved in the processes of mitosis commonly lead to chromosomal errors. For example, asbestos fibers spear cells and bind proteins that are involved in mitosis, leading to non-disjunction and other chromosomal aberrations [53]. Polycyclic aromatic hydrocarbons (PAHs) also produce chromosomal errors by causing the rapid shortening of the telomere length [54]. Tobacco smoke constituents play a prominent role in the chromosomal aberrations [55–57].

Indel Mutations

Chemical and physical carcinogens are well known to cause single and double strand breaks in DNA. Some DNA adducts, such as the alkyl-N7-deoxyguanosine formed by alkylation (or aryl-N7-deoxyguanosine by arylation) may lead to a depurinic site, which weakens the strand and easily leads to a DNA strand break. The formation of the covalent adduction on the N7-position of purine residues in DNA destabilizes the guanine-N9-1'-deoxyribose glycosidic linkage [58–60]. Single and double strand breaks also result from ionizing radiation. Two DNA strand breaks some distance apart can lead to deletions, the size of which is dependent upon the separation distance of the breaks. These carcinogen-initiated deletions may disrupt oncogenes such as RB or TP53, produce loss of heterozygosity (LOH), or result in chromosomal translocations [39, 45, 49]. Large deletions (and insertions) of one million bases or more are detectable by cytogenetics and will fall into the chromosomal aberration category.

The explanation for the formation of insertions is less clear, but may entail the same basic mechanisms as noted above for deletions, or by retrotransposition of viral or repeat sequences. The highly repeated Alu sequence, of approximately 300 nucleotides in length, is present in over a million copies in the human genome [61, 62]. Every so often, an Alu repeat sequence inserts itself in a new section of the DNA. This has been

observed as a major oncogenic mutation in retinoblastoma and numerous genetic diseases [63]. However, these events are rare and retrotransposition is probably not a prominent mechanism in cancer development.

Epigenetic Mutations

Another form of DNA mutation that has been demonstrated in recent years to be crucially involved in cancer development is the alteration of patterns of DNA 5-methyldeoxycytidine (5mdC) and/or changes in DNA chromatin modification. This type of mutation is referred to as an “epigenetic mutation,” since these changes do not alter the basic DNA coding sequence. To clarify firstly DNA 5mdC patterns, selective deoxycytidine residues in DNA are enzymatically modified by endogenous DNA methyltransferases (DNMTs) that catalyze the transfer of a methyl group from S-adenosylmethionine to the 5-position on the cytosine base, forming 5mdC [64]. In adult somatic cells, this normal DNA modification occurs almost exclusively at 5'-CpG-3' sequences and duplex DNA is maintained in a symmetrically methylated state, so that the complementary 5'-CpG-3' sequence is also methylated on the opposing strand [65, 66]. It should be noted that in stem cells, especially embryonic stem cells, a significant portion of the 5mdC also occurs in non-CpG sequences [65, 66]. The meaning of these non-CpG methylated sequences is presently unknown. Maintenance of symmetrically methylated duplex DNA ensures the mitotic heritability of these DNA 5mdC patterns in somatic cells by providing the 5mdC pattern template for each newly synthesized strand following semiconservative DNA synthesis. The level of 5mdC on a specific DNA locus or gene promoter region determines whether the sequences are expressed or not. Highly methylated promoter sequences 5' to the coding sequence of a given gene indicate that the gene is not available for transcription [67]. Today this is commonly referred to as gene hypermethylation, which is identified frequently as an epigenetic mutation in one or more tumor suppressor genes in human cancers [67, 68]. Hypomethylation of these same promoter sequences opens the gene sequences up for transcription, but does not indicate that the gene is being transcribed since additional DNA binding proteins and transcription factors are required for expression of the gene.

Not surprisingly, DNA 5mdC patterns change during tumor development [69, 70], and chemical carcinogens are known to alter DNA 5mdC patterns [71–75].

Chromatin modifications are more complex than DNA 5mdC patterns and include a number of different molecular moieties. The basic repeating unit of chromatin is the nucleosome, which is composed of 147 base pairs of DNA wrapped around a core containing two copies each of four core histone proteins H2A, H2B, H3, and H4 [76, 77]. Various molecular modifications of the N-terminal tails of these core histones occur in mammalian cells. In 1964, Allfrey et al. [78] first reported histone N-terminal tail modification, acetylation, but it was many years later before the complexity and importance of this posttranslational modification became known. As it turns out, multiple histone modifications occur, consisting of methylation [79–81], acetylation [82], phosphorylation [83], ubiquitinylation [84], and sumoylation [85]. The patterns and combinations of these modifications are referred to as the histone code [86], which influences the chromatin structure and remodeling for either active expression or silencing of genes [87]. Euchromatin (active chromatin) along with active gene expression was associated with methylation of the lysine residue in the fourth position of the N-terminal tail of histone H3 (denoted as H3K4me), while H3K9me and K3K27me formation were linked to heterochromatin (silent chromatin) and repression of gene expression [79–81]. These patterns become even more complex as each of these histone lysine residues may be mono-, di-, or tri-methylated (i.e., H3K4me¹, H3K4me², and H3K4me³) [88–91]. The importance of multiple methyl groups on lysine residues is presently only poorly understood. Histone methylation also takes place on other lysine residues on histone H3 K36 and K79, and histone H4 K20, and on arginine residues histone H3 R2, R17, and R26, and histone H4 R3.

Acetylation of the positively charged lysine residues decreases the effective ionic charge, reducing the affinity to the negatively charged phosphate backbone of DNA and creates a binding site for regulatory DNA proteins that activate transcription [92]. Essentially, histone acetylation alters the chromatin structure to a more transcriptionally supportive architecture. Acetylation of the amino group of lysine residues occurs on histone H2A K5 and K9, histone H2B K5, K11, K12,

K15 and K20, histone H3 K9, K14, K18, K23 and K56, and histone H4 K5, K8, K12 and K16 [86]. Acetylation patterns are established by highly regulated histone acetyltransferases (HATs) and histone de-acetylases (HDACs) [93]. HAT enzymes are grouped into different family classes, GNAT family, MYST family, and CBP/p300 family [94]. Some transcription factors such as yeast protein Gcn5, a putative transcriptional adaptor, are HAT enzymes [95]. HDACs catalyze the removal of acetyl groups from histones by activation of a water molecule through a zinc ion in the active site [96, 97], and are involved in formation of heterochromatin and the silencing of genes. In human cells, 18 nonredundant HDACs have been identified [98], which have been divided into classes I, II, and III [99]. Class I HDACs are localized within the nucleus while Class II shuttle between the nucleus and cytoplasm [98]. Regulation of HDACs consists of protein–protein interactions, posttranslational modifications, cellular localization, and metabolic cofactors [100].

In 1967, Gutierrez and Hnilica [101] first reported histone phosphorylation of serine residues on H3S10 and H3S28. Histone phosphorylation appears to be involved in cell cycle chromatin condensation, active transcription, apoptosis, and DNA repair mechanisms [102, 103]. Histone H3T11 and H3S28 have been identified as mitosis-specific sites for phosphorylation [104, 105]. Histone phosphorylation is also involved with activation of early response genes such as c-fos, c-myc, and c-jun. This is a stimulant-dependent H3 phosphorylation, resulting in quick transcriptional activation of a subset of immediate-early genes [106]. Although, acetylation and phosphorylation are both linked to transcription activation, it is unknown whether the two modifications work independently, in parallel, or synergistically [107, 108]. Phosphorylation also occurs on serine and threonine residues on histone H2A S1, histone H2B S14, and histone H4 S1.

Histone ubiquitinylation and sumoylation are the least understood of the chromatin modifications [84, 85]. However, these modifications may be important in DNA damage response and repair, V(D)J recombination, meiotic and spermatogenesis remodeling, and chromosomal stability [109–114].

It is becoming clear that histones are major carriers of epigenetic information. Histone modifications provide mechanisms to generate and stabilize chromatin

structures. The study of chromatin modification patterns is a very young field, which will undoubtedly become very important to the delineation of carcinogenesis. In fact, evidence is already beginning to accumulate regarding alterations in chromatin histone modifications due to exogenous chemical carcinogen exposures [75, 115].

Cytotoxicity and Cellular Turnover

Most exogenous carcinogens are cytotoxic at recognized carcinogenic doses, and probably at much lower doses to impacted cells. This concept is easy to envision when considering agents that directly damage cellular macromolecules (i.e., ionizing radiation, alkylating and arylating carcinogens). Interference with cellular functions, enzymatic processes, cytoskeletal structure, mitochondrial electron transport and oxidative phosphorylation, etc. may lead to cell senescence or death. If the DNA is damaged to the extent that successful repair is doubtful, the cell undergoes apoptosis as mediated by TP53 [39, 116]. Both senescence and apoptotic pathways are protective means by which cells may avoid accumulating DNA mutations [33].

Concentrations of chemical carcinogens and doses of physical carcinogens that produce mutagenicity in experimental studies generally also produce a significant percentage of cell lethality [117]. Whether cell death is the result of cellular damage inflicted directly by the carcinogen or from triggering apoptosis, the loss of the cell leaves a gap to be filled by cellular turnover. In tissue, this leads to increased cell division to replace these lost cells. This damage-induced cell proliferation offers increased opportunities for the accumulation of mutations during DNA synthesis and mitotic division. Cells adjacent to the vacated space will have the chance to divide to replace the missing cell. One or more of these adjacent cells might harbor a mutation that confers a growth advantage enabling it to dominate over its neighbors and progress through cell division. Thus, the potential for clonal expansion and the continued accumulation of oncogenic mutations is encouraged by the impact of carcinogens on tissues.

The opportunity or encouragement of carcinogen exposed cells to undergo cell division provides yet another enhancement for mutation occurrence. Many of

the cells that have the opportunity to divide have also received a carcinogen-induced damaging event, which if not repaired prior to DNA synthesis may lead to a fixed mutation [117]. For example, the presence of an alkyl-O⁶dG adduct during DNA synthesis will result in the formation of a G to T mismatch approximately 50% of the time. Alkyl-O⁶dG base pairs equally efficiently with either a thymidine or a cytidine residue [60, 118, 119]. Different DNA adducts and forms of damage will result in different mutational events following DNA synthesis. But the cytotoxic action of these carcinogens couples with the DNA damaging activity to induce mutations.

When carcinogen induced cell death involves a tissue stem cell, the dead cell is replaced by another stem cell. Most of cancer arises from tissue stem cells. The accumulation of mutations in a transit-amplifying progenitor cell will usually result in no consequence as the developing cancer cell will be terminally lost either by terminal differentiation and cell death or by being sloughed off the epithelial layer [34]. Stem cells have the capability of surviving long enough to accumulate the necessary mutations to become cancerous. Even in rapidly turning over tissues, stem cells appear to be long lived, with estimated lifetime division numbers totaling in the hundreds and higher [120]. But these stem cells also enjoy a slow or delayed cell cycle, dividing intermittently (only once every 7 or 8 days), and often sitting quiescent and in a nonproliferative or reserve stem cell state [121–123]. Stem cell division is generally considered to be asymmetric, giving rise to both a replacement stem cell and a progenitor cell that will undergo numerous symmetrical cell divisions and then terminally differentiate [120, 124–126]. This asymmetric cell division appears to also include the retention of the original DNA strand by the daughter stem cell replacement. Or stated differently, the original template DNA strand of the mother stem cell selectively segregates to the daughter stem cell replacement, while the newer DNA strand segregates to the transit-amplifying daughter cell. This is the Cairns' "immortal stranding" hypothesis [127], which appears to have experimental validity [124, 126]. DNA synthesis in human cells accumulates one to three mutations every time the genome is copied [31, 32]. "Immortal stranding" enables the stem cell to minimize the accumulation of DNA mutations, and thus limit the

number of mutations that are passed on to the transit-amplifying differentiating epithelial cells.

Adult stem cells are very sensitive to DNA damage, and rather than perform DNA repair, stem cells appear to undergo apoptosis in response to external exposure to DNA damaging agents [128]. Discarding the cell rather than risk faulty repair ensures that the accumulation of DNA mutations in stem cells does not occur. However, the loss of one or more stem cells due to external toxic exposure(s) (UV light, ionizing radiation, arylating or alkylating chemicals, or inflammation or infection produced ROS), requires the replacement of these stem cells. Stem cell replacement requires a symmetrical cell division to give rise to two daughter stem cells, which will produce one “aged” DNA template strand regardless of “immortal stranding.” This process increases the number of mutations arising in the stem cells and increases the risk of disease. Alternatively, a transit-amplifying cell may undergo dedifferentiation, characteristic of mesenchymal-epithelial transition [129]. Unfortunately, this dedifferentiation process again produces a stem cell without the benefit of “immortal stranding” and increases the risk of mutations and disease developing in the tissue stem cell population.

Genomic Instability

Genomic instability is the loss of genomic stability and the increased probability for accruing DNA mutations. Restated, genomic instability is the inability of cells to maintain the integrity of their own genomic inheritance.

It is generally accepted that most cancers express one or more forms of genomic instability. However, among the scientific community there is still controversy as to whether the loss of genomic stability is required for the development of cancer. One argument states that genomic instability is not necessary for the accrual of a sufficient number of oncogenic mutations to reach a malignancy [130, 131]. The other argument states that human cells would not have the opportunity to accumulate a sufficient allotment of driver mutations to achieve a cancerous cell without an additional navigating mechanism(s) [132]. The statistical probability during a normal human life span has been proposed to be too low based on theoretical cell division numbers alone. However, many oncogenic mutations, such as in “caretaker” genes, result in the enhanced probability of

initiated cells achieving more mutations and progressing along the road of tumor development [31, 40, 42, 133]. Caretakers are genes that code for cellular proteins and factors that protect the genome from mutation. Disruption of the finely balanced cellular network of monitoring and repair systems will lead to a loss of fidelity of these genome protection systems, increases in macromolecular damage, increased longevity of DNA damage, faulty repair, and an increased accumulation of DNA mutations. For example, a mutation that disrupts the normal function of a mismatch DNA repair gene leads to upward of a 100-fold increased rate of cellular DNA mutation [40].

Genomic instability is a hallmark of cancer and one of the trademarks of a number of inherited genetic diseases that predispose individuals to cancer [131] (see Table 1). By today’s laboratory techniques, genomic instability manifests as chromatin instability (CIN), microsatellite instability (MSI) (sometimes referred to as MIN), nucleotide instability (NIS) (also called single base instability), and epigenetic instability (EIS) (see Table 2).

Chromosomal Instability (CIN)

The inability of the cell to properly repair single or double strand DNA breaks, perform appropriate recombination activities in somatic cells, or correctly perform the mitotic spindle functions during mitosis will lead to the development and accumulation of chromosomal aberrations, which is the manifestation of CIN [33, 134]. Cells expressing CIN miss-segregate chromosomes 10–100 times more frequently than chromosomally stable diploid cancer cells [135]. CIN may promote tumor evolution by enabling the clonal expansion of cells with proliferative advantages and/or metastatic potential [136–138].

Chemical and physical carcinogens can interfere with the cytoskeleton, mitotic spindle apparatus, or the various proteins and enzymes involved in mitosis. The impact of such carcinogenic damage is the increased probability of chromosomal mutations. As noted above, asbestos produces aneuploidy in skewed cells by interfering with the cytoskeleton and mitotic proteins [53]. Tobacco smoke PAHs, BPDE, and 4-aminobiphenyl enhance telomere shortening and produce CIN in vivo and in vitro studies [54–57].

Microsatellite Instability (MSI)

Microsatellites are relatively short sequences composed of tandemly repeated units of one to six bases in length [139]. The repeated unit of sequence may be only a mono-nucleotide repeat (e.g., AAAAAA...), di-nucleotide repeat (e.g., CACACACACACA...), tri-nucleotide repeat, etc., although microsatellites of imperfect and compound sequences also exist in the human genome (e.g., Bat25 and Bat26) [140]. The most common class of microsatellites in the human genome are composed of various tandemly repeated lengths of (CA)_n. Microsatellites are widely prevalent throughout the human genome, occurring more frequently in non-coding regions on all chromosomes, and act as gene promoters, as target sites for recombination, and as binding sites for DNA topoisomerases [139]. The length of a specific microsatellite is an inherited trait and therefore the same in all tissues. Insertion or deletion of one or more repeat units within a microsatellite occurs on rare occasion by slippage during DNA replication, resulting in a longer or shorter microsatellite sequence, respectively. These polymerase slippage errors are normally repaired by a post-replication mismatch repair system, and the presence of a microsatellite variant generally results from the lack of fidelity in the mismatch repair mechanism [141–143]. MSI is denoted by the presence of microsatellite variants of different lengths in tumors compared to normal tissue, and is a hallmark of mismatch repair deficient familial and sporadic cancers [144, 145].

In theory, a carcinogenic agent that interferes with DNA polymerases and/or the mismatch repair system would lead to MSI. For many types of cancer, MSI is the one genomic instability that is consistently acquired late in the tumorigenic process [139]. The major exception to this rule is the hereditary nonpolyposis colorectal cancer (HNPCC) syndrome, for which individuals inherit an error in the mismatch repair system and are at risk of developing colon cancer [146, 147]. However, airborne particulates (PM_{2.5}) have been associated with the development of MSI in human lung epithelial cells [148, 149]. Cancers induced by betel-quid chewing and chronic occupational exposure to chromium have been found to express MSI significantly more frequently than noncarcinogen-associated tumors [150, 151].

Nucleotide Instability (NIS)

Nucleotide instability may be one of the more easily induced instabilities by either exogenous or endogenous influences. Increases in the frequencies of various forms of DNA nucleotide damage and/or loss of cellular DNA damage surveillance and repair capacity, or the perturbation of the DNA polymerases or synthesis complex machinery result in increases in single base mutations [19, 46–48, 132]. With the possible exception of chromosomal translocations in leukemias and lymphomas, NIS is probably the most common genomic instability expressed in human cancer [47, 49, 50]. NIS may also be predominant in the earliest mechanisms of carcinogenesis as roughly 85% of the activating and deactivating mutations in proto-oncogenes and tumor suppressor genes, respectively, are single base substitution mutations [44–48]. Much of the NIS data in the literature has been accumulated from clonal expansion of mutations in tumor tissue, that is, TP53 mutation database and retinoblastoma gene mutations [49, 50]. The analysis of NIS has been hampered by the requirement of exceedingly sensitive techniques, and the limited viable ultrasensitive approaches are labor intensive [47]. However, with the present availability of the ultra-sensitive methods, the presence of NIS in non-tumor tissues and the induction of NIS by exogenous environmental carcinogens are beginning to unfold [48, 152].

Epigenetic Instability (EIS)

Although any alteration from normal of DNA 5mdC and/or chromatin modification patterns constitutes epigenetic instability, the majority of the scientific literature has focused on hypermethylation, that is, the transcriptional suppression of important anticancer genes by high levels of 5mdC present in the gene promoter sequences [67, 68]. In human cancer, a hypermethylation phenotype, referred to as CIMP (CpG island methylator phenotype), has been proposed [153]. The CIMP phenotype may have clinical significance, but the cause or mechanism underlying the CIMP tumor phenotype has yet to be identified [154–156]. Regardless of the gene specific hypermethylation ascertainment biased approach, cancer cells are globally hypomethylated and harbor 20–60% less genomic 5mdC than their normal counterpart. The hypomethylated sequences in cancer cells

may be mainly due to the loss of 5mdC levels in repetitive sequences, which are normally heavily methylated and comprise 20–30% of the human genome [157]. For at least some cancers, the level of global hypomethylation is associated with progression, becoming progressively more hypomethylated from normal to benign, to tumor and metastasis [158–160]. Indeed, transient global hypomethylation is inducible by either chemical carcinogens or hypoxia [71, 72, 161]. Interestingly, increased NO production that may occur in inflammatory disease, or in response to hypoxia, reduces S-adenosylmethionine levels by inactivating methionine adenosyltransferase, and produces hypomethylation [162]. Dietary deficiencies in either folic acid or methyl donor compounds, that is, methionine, also cause hypomethylation [6–8]. Hypomethylation of gene-specific sequences in cancer cells leads to ectopic gene expression, such as neuron-specific γ -synuclein expression in gastric cancers [163]. The degree of hypomethylation of the XIST gene correlates with prostate cancer aggressiveness [164].

There may be several mechanisms at work in the alterations of epigenetic signals during carcinogenesis as these patterns are intimately involved in all biological processes, that is, genomic imprinting, gametogenesis, embryogenesis, differentiation, aging, and carcinogenesis. As noted above, dietary deficits are known to alter DNA 5mdC levels and patterns [6–8]. Alkylation, arylation and other forms of DNA damage induced by carcinogens have been clearly shown to disrupt DNA 5mdC patterns [71, 72]. DNA epigenetic patterns may also be impacted by many other environmental factors, including tobacco use, drug use, physical exercise, and health status as DNA 5mdC patterns diverge with age in monozygotic twin pairs [165].

Chromatin modification patterns are well known to be involved in carcinogenesis and the development of cancer [166, 167]. However, this field is still developing, and the data regarding exogenous agent-induced changes in histone modifications are presently limited. Several in vitro tissue culture studies have been performed that suggest toxicants induced changes in selective histone residue methylation or acetylation (see review [75]), but the interpretation of these specific histone modification changes amid the complex arrays of epigenetic patterns and extrapolation to in vivo will require further analyses.

To summarize genomic instability, there are a large number of genes involved in the maintenance of the genome. Many of these genes, such as the DNA repair genes, have been referred to as caretaker genes for their role in genomic stability, while others such as the RB and TP53 genes have been denoted as gatekeeper genes due to their involvement in cell cycle control and the pathway to cancer development [168]. Each of these instabilities enhances the probability of accumulating oncogenic mutations and progression along the road of tumor development, the acquisition of invasive or metastatic capability, and/or resistance to chemotherapeutic regimens. This last effect has also been observed in non-chemotherapeutic carcinogens, whereby cells exhibiting selective genomic instabilities are resistant to the cytotoxic effects of some carcinogens. Cells expressing CIN were found to be resistant to the lethal effects of PhIP but not to the methylating carcinogen, *N*-methyl-*N'*-nitro-*N*-nitrosoquandine (MNNG), while cells expressing MSI were resistant to MNNG but not to PhIP [169]. This provides yet another important mechanism by which chemical carcinogens may select for mutant cells and promote their clonal expansion by exerting cytotoxic effects on neighboring cells in the tissue.

Susceptible Individuals

Based on the above discussion of genomic instabilities, it is not surprising that some individuals in the human population have inherited genetic errors that render them more prone to developing cancer [47, 131] (see Table 1). For a portion of these susceptible individuals, the risk of cancer is quite high. For example, individuals who have inherited a mutation in one of the mismatch repair genes (i.e., MSH2, MLH1, MSH3, MSH6, PMS1, and PMS2) are carriers of hereditary nonpolyposis colorectal cancer (HNPCC) syndrome and are at high risk of colorectal carcinoma [139, 146, 147]. Inheritance of any one of several DNA repair gene mutations confers a genetic disease (i.e., xeroderma pigmentosum, ataxia telangiectasia, Fanconi anemia, Cockayne syndrome, and others), which is associated with adverse clinical symptoms including a substantially elevated risk of cancer. Individuals harboring a germ line mutation in the TP53 gene, referred to as the guardian of the genome gene, generally present as

a Li-Fraumeni Cancer Family Syndrome and are at a substantially elevated risk of multiple types of cancers during their lifetime [170]. More than 100 Mendelian diseases that confer predisposing cancer syndromes have been recorded in “McKusick’s Online Mendelian Inheritance in Man” (OMIM), each with varying levels of cancer risk [171].

There are however heritable mutations that convey more subtle increases in cancer risk, especially when environmental exposures to exogenous carcinogens are involved [47, 172]. Polymorphisms in genes involved in the deactivation and elimination of toxic metabolites and carcinogens may confer to the individual a decreased ability to remove offending toxicants and/or reactive electrophiles before they can cause DNA damage, which provides a slightly increased probability of developing cancer. Examples of these susceptibility traits are found in polymorphisms among the phase I and phase II biotransformation enzymes, such as cytochrome P450 and glutathione-S-transferase genes [172–174]. These individuals are at an increased risk of cancer from exposures to polycyclic aromatic hydrocarbon (PAH) carcinogens commonly found in tobacco smoke and other combustion products [173, 175]. Depending upon the metabolic pathway used to activate, deactivate, and/or eliminate carcinogens, gene polymorphisms may render individuals more susceptible or less susceptible to cancer from exposure to selective carcinogenic agents.

Dose and Complex Dose

Humans are exposed to a plethora of toxic and carcinogenic materials in the environment. Exposure to complex toxic mixtures poses a higher risk of detrimental health effects than that represented by the toxicity of each separate individual constituent. When an individual or animal is exposed to a complex mixture of carcinogenic agents, the dose required to produce significant health problems and tumor development will be lower than that for each separate individual constituent. In toxicology, it is well recognized that exposures to significant levels of complex mixtures of toxic compounds results in increased detrimental health effects. This is also true for carcinogenic agents. The combined toxic effect of separate toxicants is generally predicted to be at

a minimum additive, unless specific information exists to demonstrate a synergistic (above additive) or inhibitory (below additive) action. Many of these carcinogens target the same organs or tissues, and induce similar toxic effects in humans and animals. For example, alcohol use, foodborne aflatoxins and hepatitis B (or C) all adversely impact the liver and promote the development of hepatocellular carcinoma. Similarly, asbestos, arsenic, and tobacco smoke drive the development of lung cancer. For each of these examples, some differences in agent-specific direct DNA damage may exist, but similar mechanisms of cellular damage, enhanced cellular turnover, inflammatory responses, and ROS and RNS species involvement contribute to the development of cancer.

Although people are not generally exposed to an acute carcinogenic dose for each one of these environmental carcinogens, they may receive an accumulated dose sufficient to induce cancer due to long-term chronic low dose exposure. Damage to the skin from lifetime exposure to UV radiation from sunlight is one of the more obvious examples. However, the same may be true for a number of environmental chemical carcinogens, especially if one has been chronically exposed to tobacco smoke, whether as a smoker or nonsmoker. What is known regarding what constitutes a carcinogenic dose is based mostly on laboratory studies involving a single agent, not a complex mixture. It should note, however, that tobacco smoke is a complex mixture, and epidemiological studies have provided estimates of carcinogenic risk based on the number of cigarettes smoked per day and the number of pack years of smoking [176, 177]. Estimates of carcinogenic risk are generally derived from the carcinogenic dose of individual agents. The combined effects of exposures to multiple environmental carcinogens must certainly lower the carcinogenic dose threshold for individual carcinogens. Thus the chronic long-term low-dose exposure to a multitude of carcinogenic agents in our food, air, water, from sunlight, x-rays, and biological pathogens, enhances the risk of cancer as our bodies reach old age. Whether cancer is inherent with life or not, moderating one’s exposure by avoiding habitual exposures, that is, tobacco smoke and alcohol abuse, and maintaining a healthy lifestyle significantly lowers one’s lifetime risk of cancer.

Environmental Carcinogens

To reiterate, the rate of tumor development is increased sometimes many orders of magnitude above the spontaneous background rate by exposure to exogenous carcinogens. The timing of the onset of cancer is due to the dose and duration of carcinogen exposure and not to either the chronological age or the age that exposure began [38, 39]. Regardless of whether a cancer develops from carcinogen exposure(s) or endogenous processes, or the combination of exogenous and endogenous attacks (the most likely explanation), the accumulation of oncogenic driver mutations and a large number of cell divisions are required.

Separating the effects of exogenous versus endogenous genotoxic damage and cancer induction may not be an easy undertaking. More than likely, a tumor originating only from exogenous carcinogen exposure is rare if nonexistent. However, the impact of exogenous exposures to genotoxic agents is easily demonstrated by the significant increases in lung cancer of smokers [176], and by the patterns of oncogenic single base substitution mutations in the TP53 gene in lung cancers and breast carcinomas from smokers. Lung and breast tumors arising in smokers are twice and ten times, respectively, as likely to harbor a TP53 gene driver mutation of the G to T transversion type than similar tumors arising in nonsmokers [178, 179]. Benzo(a)pyrene and other PAHs are known to target deoxyguanosine in the most commonly mutated sites (codons 157, 248, and 273) observed in lung tumors, and to produce G to T transversion mutations in these same base sites [180–182]. Similar findings have reported for aflatoxin B1 adduction and G to T mutation of codon 249 of TP53 [183–185]. Sunlight (UV light) produces signature CC to TT mutations in TP53 and other DNA sequences that have been identified in sun exposed skin, which is well correlated with the occurrence of skin cancers [186–188]. Unfortunately, few carcinogens presently provide such clearly apparent tumor-associated agent-induced genotoxic damage. Signature diseases that aid in identifying the exogenous impact of toxicant exposure are also limited to only a few carcinogens, for which asbestos and mesothelioma are the best known examples. Many other cancers have

multiple etiological causes, and delineation of specific carcinogen exposures in the causation is not easily identified (see Table 1).

The following are brief descriptions of the genotoxic processes and carcinogenic mechanisms for a few examples of common environmental carcinogens, which are provided here to clarify the discussions of the impact of these exposures on the development of human cancer. The discussions are not intended to be fully comprehensive. Extensive literature is available of each of these agents and numerous references are provided for the reader.

Alcohol

Chronic alcohol consumption is associated with oral cavity, pharynx, larynx, esophageal, liver, and colorectal cancer, and breast cancer in women [12, 13, 16]. Alcohol use is also suspected to be involved in the etiology of pancreatic and lung cancers. The development of alcohol-induced cirrhosis of the liver is a well-known etiology for hepatocellular carcinoma development [16], and the risk of hepatocellular carcinoma increases with the synergistic impact of tobacco smoke and alcohol, or of hepatitis (B or C) and alcohol [12]. However, pure ethanol itself is not genotoxic, but its primary metabolite acetaldehyde is mutagenic [189, 190].

The present theories regarding the carcinogenic mechanisms of alcohol are based on DNA damaging events, tissue inflammation and perturbation of nutrition and physiological functions [13, 14]. Acetaldehyde forms DNA adducts [189, 190], and damages hepatocytes and tissue architecture, which leads to increased cellular proliferation [191]. Alcohol induces CYP2E1, whose gene product (P450 2E1) adds to the metabolism of alcohol to acetaldehyde as well as activating tobacco carcinogens (i.e., nitrosamines) [12]. Alcohol also induces ROS & RNS [192], which leads to lipid peroxidation and further DNA damage from reactive electrophilic products of the lipid peroxidation reactions [193]. Alcohol intake interferes with folate metabolism [194] and folate deficiency increases the risk of colorectal carcinoma by two to five times [195]. Alcohol may also alter the absorption and metabolism of vitamins, such as vitamin B₁₂, B₆, and retinoic acid, and often results in malnourishment in the

alcoholic [13, 14, 194]. Thus chronic alcohol use utilizes at least three different avenues of carcinogenesis; DNA damage and mutation, inflammatory processes that promote proliferation, and vitamin deficiencies that are known to perturb DNA epigenetic patterns. It is not surprising that alcohol works synergistically with other carcinogens such as tobacco smoke to increase the occurrence of cancers [12].

Sunlight

UV light from the sun is probably the most important naturally occurring environmental carcinogen to which humans are exposed. Chronic exposure to sunlight and repeated sunburns are strongly correlated with nonmelanoma skin cancers (mainly squamous cell carcinoma and basal cell carcinoma), especially on the head, neck, forearms, and hands, which clothing generally leaves uncovered [196]. Skin cancer is one of the more frequent cancers in aging populations, and is the most common cancer in the USA [9, 10]. The lifetime risk of developing skin cancer in the USA is one in five. Fortunately, most skin cancers are amenable to early detection and a high cure rate. Melanoma is the exception, leading to a lethal disease course in greater than 10% of cases.

UV light is comprised of electromagnetic radiation wavelengths below 400 nm, and has been broken down to vacuum UV (100–200 nm), UVC (200–290 nm), UVB (290–320 nm), and UVA (320–400 nm). [Note that some authors refer to the wavelength ranges for UVB and UVA as 280–315 nm and 315–400 nm, respectively (e.g., Ridley et al. 2009 [197]).] The upper layers of the Earth's atmosphere filter out much of the harmful UV shorter wavelength, although this will become less effective and a potentially major health issue as the ozone layer thins over parts of the planet. The remaining UVA and UVB radiation reaches the ground. Both UVA and UVB damage cells and macromolecules in cells. UVB provides the major DNA mutational impact producing signature cyclobutane pyrimidine dimers and pyrimidine 6–4 pyrimidone photoproducts, which if not repaired, result in signature CC to TT and C to T mutations, respectively [186, 187, 198]. Cyclobutane pyrimidine dimers are responsible for approximately 80% of the UV light signature

mutations, since the 6–4 photoproducts are repaired much more efficiently than the cyclobutane pyrimidine dimers [188, 197]. Cyclobutane pyrimidine dimers are only slowly repaired by a transcription-coupled repair system. These UV signature mutations are known to occur in at least one important tumor suppressor gene for human cancer, the TP53 gene [186]. UV-induced TP53 mutations probably represent an early step in the skin cancer carcinogenesis pathway for basal cell carcinoma and squamous cell carcinoma [186, 199].

UVA radiation also contributes to the carcinogenic impact of sunlight exposure. UVA radiation leads to DNA damage and mutations, albeit at more than a 1,000-fold lower extent than UVB. However, UVA penetrates more deeply into the skin than UVB and produces UV signature mutations in the basal layer, while the majority of UVB produces signature mutations in suprabasal keratinocytes [200]. UVA also initiates substantial cellular macromolecular damage, ROS, and lipid peroxidation [197].

UV light exposure may enhance the development of skin tumors both by inducing mutations and by interfering with the turnover of keratinocytes [201–203]. UV irradiation induces apoptosis in severely damaged keratinocytes. However, insulin-like growth factor-1 (IGF-1) produced by dermal fibroblasts opposes UV-induced apoptosis in keratinocytes and induces replicative senescence by NF κ B and TP53 signalling [201–204]. UV damaged keratinocytes that survive apoptosis still retain the capacity to divide unless held in check by IGF-1. Aging skin fibroblasts produce significantly less IGF-1, leading to the increased potential of damaged and mutant keratinocytes replicating. This may explain the increased incidence of nonmelanoma skin cancer in geriatric patients [205].

The combined effects of UVB and UVA to multiple cells in the epidermal layer result in localized tissue inflammation, which contributes to the melee of oxidative reactions and includes reactive nitrogen species [25, 27, 43]. The setting of cellular damage and inflammation results in a tissue microenvironment that encourages cellular turnover, promotes clonal expansion of mutated cells that leads to the accumulation of oncogenic mutations and ultimately to skin cancer development.

Tobacco Smoke

Tobacco smoke is the most frequent cause of lung cancer and probably the most important anthropogenic carcinogen of the last century. Whether by life-style choice of smoking or by exposure to side-stream smoke for the nonsmoker, tobacco smoke has been a primary etiological factor in lung cancer in the USA [176]. On average, lung cancer develops in cigarette smokers after 50 pack-years [177]. Tobacco smoke contains a large complex mixture of carcinogens, including many that may both initiate and promote the development of cancer. Cigarette smoke contains between 3,000 and 4,000 toxic compounds, more than 300 of which are known or suspected carcinogens, including the majority of the IARC Group I chemical carcinogens discussed in this chapter [206] (see Table 1). Many of these toxic and carcinogenic components of tobacco smoke are readily absorbed systemically, and smoking is well associated with diseases such as cardiovascular disease and cancers in addition to lung [176, 177, 207]. Tobacco smoking is believed to be an important etiological factor in many cases of oral cancers, esophageal, bladder, pancreatic, kidney, and possibly breast and cervical cancers [208].

Many of the carcinogens in tobacco smoke alkylate or arylate DNA, producing promutagenic lesions. For example, benzo(a)pyrene is activated by the host cells' cytochrome P450 1A1 enzyme activity to the ultimate carcinogen, benzo(a)pyrene-7,8-dihydrodiol-9,10-epoxide (BPDE) form, which reacts with nucleophiles including the exocyclic amino group on deoxyguanine of DNA [209]. The BPDE-N²dG adduct is a promutagenic lesion that leads most frequently to a G to T transversion mutation. Numerous other tobacco smoke carcinogenic PAHs, heterocyclic amines, and nitrosamines, produce DNA adducts that can lead to primarily G to T transversions. As noted above, G to T transversions are the most common single base mutations in lung tumors [178, 179]. A few of these carcinogenic nitrosamines are tobacco smoke specific, that is, 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) and N'-nitrosonornicotine (NNN) [207, 210]. The list of carcinogens in tobacco smoke is extensive and ensures multiple types of DNA damage and adverse impacts on

other cellular macromolecules, enzymes, and structural components is complex [208]. It has been estimated that a developing preneoplastic clone of lung cells accumulates one DNA mutation for every 15 cigarettes smoked [211]. In addition to damaging DNA, most of these tobacco smoke toxicants also promote cellular turnover and inflammatory processes in lung tissue, providing another avenue of ROS and RNS species and altered microenvironments for clonal expansion of mutant cells [25, 43]. Particulates from tobacco smoke interfere with pulmonary function independent of the plethora of toxic chemicals, recovery and repair following exposure [148, 149]. In spite of all these genome destabilizing and tumor-promoting activities, many years of smoking history are generally required before a tumor develops, which is a testimony to the resilience of the human body to the bombardment of such a powerful combination of the carcinogens and toxic agents present in tobacco smoke.

Future Directions

It would be very difficult if not impossible for life on this planet to completely avoid exposure to environmental carcinogens. One can purposely avoid alcohol, tobacco smoke, and many occupational exposures, but the presence of carcinogenic substances in our food, air, water, and sunlight coupled with health, nutritional issues, and biological carcinogenic agents make avoiding all exposure to carcinogens very difficult. The extensive years of educating the public regarding the hazards of smoking, along with the banning of smoking in public buildings and other establishments has had a dramatic effect on the reduction in lung cancer and probably other cancers in recent years [176, 177]. Increased safety precautions and worker awareness to reduce occupational exposures remain a very important component in reducing cancer risk. Vaccinations against biological carcinogens, such as hepatitis B and C, may be important as well, and perhaps more important depending upon where one travels in the world. As people live longer and healthier lives, reductions in carcinogenic contaminants in our food, air, and water appears to be the only other viable avenue that society as a whole can take to diminish cancer risk. Individuals remain responsible for their

own nutritional health, exercise, and personal habits in avoiding carcinogenic exposures.

Fortunately, our food also contains anticarcinogens [212–214]. For example, the isothiocyanates in broccoli are suggested to aid in diminishing the carcinogenic impacts of a number of chemical carcinogens [215]. This is a relatively young, but growing field of research that has the potential to provide viable strategies for long-term reductions in cancer risks. Individuals may only have to monitor their dietary intake of selected vegetables and herbs to enjoy significantly reduced cancer risk. However, this may work for reducing some types of cancer and not others.

The relative contribution of endogenous versus exogenous causes of human cancer is still a matter of some debate. Since the carcinogenic impacts of endogenous and exogenous are completely intertwined, separating out individual contributions is very difficult. Only in selective cases where signature diseases and/or signature biomarkers such as specific DNA adducts and/or mutations are identifiable can the contributions be estimated in rough manner [116]. This is a problem that is likely to remain for many years to come. Learning and understanding many of the xenobiotics in our food and environment to which humans are presently exposed continues to be a prominent field of safety research. Add in the large influx of synthetic chemical compounds joining our environmental melee each year, many of which may contribute to the carcinogenic impacts, and identifying all carcinogens in our environment becomes a rather daunting undertaking. For example, the explosion in designed nanoparticles for industrial and medical purposes, introduces potentially genotoxic agents into the realm of human exposure. At least some synthetic nanomaterials have thus far been found to cause chromosomal aberrations, DNA breaks, point mutations, and oxidative DNA damage [216]. Control of occupational exposures and environmental releases of these new genotoxic synthetic compounds will be important in opposing additions to the causes of environmental carcinogenesis.

The lessons learned from the classical examples of environmental carcinogenesis (sunlight and skin cancer, tobacco smoke and lung cancer, and aflatoxin, hepatitis and liver cancer) have guided research studies and advances in understanding cancer development.

Continued advances in understanding the tissue, cellular and molecular processes of carcinogenesis will enable the design of strategies to further diminish cancer incidence and reduce risks throughout life.

Bibliography

1. Ferguson LR, Philpott M (2008) Nutrition and mutagenesis. *Annu Rev Nutr* 28:313–329
2. Dybing E, O'Brien J, Renwick AG, Sanner T (2008) Risk assessment of dietary exposures to compounds that are genotoxic and carcinogenic—an overview. *Toxicol Lett* 180(2):110–117
3. Hashimoto K, Higuchi M, Makino B, Sakakibara I, Kubo M, Komatsu Y, Maruno M, Okada M (1999) Quantitative analysis of aristolochic acids, toxic compounds, contained in some medicinal plants. *J Ethnopharmacol* 64(2):185–189
4. Nortier JL, Martinez MC, Schmeiser HH, Arlt VM, Bieler CA, Petein M, Depierreux MF, De Pauw L, Abramowicz D, Vereerstraeten P, Vanherweghem JL (2000) Urothelial carcinoma associated with the use of a Chinese herb (*Aristolochia fangchi*). *N Engl J Med* 342(23):1686–1692
5. Lai MN, Wang SM, Chen PC, Chen YY, Wang JD (2010) Population-based case-control study of Chinese herbal products containing aristolochic acid and urinary tract cancer risk. *J Natl Cancer Inst* 102(3):179–186
6. Christman JK, Sheikhejad G, Dizik M, Abileah S, Wainfan E (1993) Reversibility of changes in nucleic acid methylation and gene expression induced in rat liver by severe dietary methyl deficiency. *Carcinogenesis* 14(4):551–557
7. Pufulete M, Al-Ghnam R, Rennie JA, Appleby P, Harris N, Gout S, Emery PW, Sanders TA (2005) Influence of folate status on genomic DNA methylation in colonic mucosa of subjects without colorectal adenoma or cancer. *Br J Cancer* 92(5):838–842
8. Pogribny IP, Ross SA, Wise C, Pogribna M, Jones EA, Tryndyak VP, James SJ, Dragan YP, Poirier LA (2006) Irreversible global DNA hypomethylation as a key step in hepatocarcinogenesis induced by dietary methyl deficiency. *Mutat Res* 593(1–2):80–87
9. Robinson JK (2005) Sun exposure, sun protection, and vitamin D. *JAMA* 294(12):1541–1543
10. Rogers HW, Weinstock MA, Harris AR, Hinckley MR, Feldman SR, Fleischer AB, Coldiron BM (2010) Incidence estimate of nonmelanoma skin cancer in the United States, 2006. *Arch Dermatol* 146(3):283–287
11. Kao SC, Reid G, Lee K, Vardy J, Clarke S, van Zandwijk N (2010) Malignant mesothelioma. *Intern Med J* 40(11):742–750
12. Boffetta P, Hashibe M (2006) Alcohol and cancer. *Lancet Oncol* 7(2):149–156
13. Seitz HK, Stickel F (2007) Molecular mechanisms of alcohol-mediated carcinogenesis. *Nat Rev Cancer* 7(8):599–612
14. Pelucchi C, Gallus S, Garavello W, Bosetti C, La Vecchia C (2008) Alcohol and tobacco use, and cancer risk for upper aerodigestive tract and liver. *Eur J Cancer Prev* 17(4):340–344

15. Hecht SS, Kassie F, Hatsukami DK (2009) Chemoprevention of lung carcinogenesis in addicted smokers and ex-smokers. *Nat Rev Cancer* 9(7):476–488
16. McKillop IH, Schrum LW (2009) Role of alcohol in liver carcinogenesis. *Semin Liver Dis* 29(2):222–232
17. He S, Nakada D, Morrison SJ (2009) Mechanisms of stem cell self-renewal. *Annu Rev Cell Dev Biol* 25:377–406
18. United States Cancer Statistics (USCS) (1999–2006) Cancer Incidence and Mortality Data, National Program of Cancer Registries (NPCR), Center for Disease Control and Prevention, Department of Health and Human Services, USA. <http://apps.nccdc.cdc.gov/uscs/index.aspx>. Accessed March 2010
19. Bielas JH, Loeb LA (2005) Mutator phenotype in cancer: timing and perspectives. *Environ Mol Mutagen* 45(2–3):206–213
20. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. *Nature* 458(7239):719–724
21. Ames BN (1989) Endogenous oxidative DNA damage, aging, and cancer. *Free Radic Res Commun* 7(3–6):121–128
22. Loft S, Poulsen HE (1998) Estimation of oxidative DNA damage in man from urinary excretion of repair products. *Acta Biochim Pol* 45(1):133–144
23. Shigenaga MK, Gimeno CJ, Ames BN (1989) Urinary 8-hydroxy-2'-deoxyguanosine as a biological marker of in vivo oxidative DNA damage. *Proc Natl Acad Sci USA* 86(24):9697–9701
24. Cooke MS, Evans MD, Dizdaroglu M, Lunec J (2003) Oxidative DNA damage: mechanisms, mutation, and disease. *FASEB J* 17(10):1195–1214
25. Hussain SP, Harris CC (2007) Inflammation and cancer: an ancient link with novel potentials. *Int J Cancer* 121(11):2373–2380
26. Dahlmann HA, Vaidyanathan VG, Sturla SJ (2009) Investigating the biochemical impact of DNA damage with structure-based probes: abasic sites, photodimers, alkylation adducts, and oxidative lesions. *Biochemistry* 48(40):9347–9359
27. Sedelnikova OA, Redon CE, Dickey JS, Nakamura AJ, Georgakilas AG, Bonner WM (2010) Role of oxidatively induced DNA lesions in human pathogenesis. *Mutat Res*.
28. Yuan J, Narayanan L, Rockwell S, Glazer PM (2000) Diminished DNA repair and elevated mutagenesis in mammalian cells exposed to hypoxia and low pH. *Cancer Res* 60(16):4372–4376
29. Camenisch U, Naegeli H (2009) Role of DNA repair in the protection against genotoxic stress. *EXS* 99:111–150
30. Goldsby RE, Lawrence NA, Hays LE, Olmsted EA, Chen X, Singh M, Preston BD (2001) Defective DNA polymerase- δ proof-reading causes cancer susceptibility in mice. *Nat Med* 7(6):638–639
31. Loeb LA, Bielas JH, Beckman RA (2008) Cancers exhibit a mutator phenotype: clinical implications. *Cancer Res* 68(10):3551–3557, discussion 3557
32. Zegerman P, Diffley JF (2009) DNA replication as a target of the DNA damage checkpoint. *DNA Repair (Amst)* 8(9):1077–1088
33. Hanahan D, Weinberg RA (2000) The hallmarks of cancer. *Cell* 100(1):57–70
34. van der Flier LG, Clevers H (2009) Stem cells, self-renewal, and differentiation in the intestinal epithelium. *Annu Rev Physiol* 71:241–260
35. Weinberg RA (2008) Mechanisms of malignant progression. *Carcinogenesis* 29(6):1092–1095
36. Weinberg RA (2008) Coevolution in the tumor microenvironment. *Nat Genet* 40(5):494–495
37. Chatten C, Bathe OF (2010) Immunoregulatory cells of the tumor microenvironment. *Front Biosci* 15:291–308
38. Peto J (2001) Cancer epidemiology in the last century and the next decade. *Nature* 411(6835):390–395
39. Loeb LA, Harris CC (2008) Advances in chemical carcinogenesis: a historical review and prospective. *Cancer Res* 68(17):6863–6872
40. Jiricny J (2006) The multifaceted mismatch-repair system. *Nat Rev Mol Cell Biol* 7(5):335–346
41. McCulloch SD, Kunkel TA (2008) The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res* 18(1):148–161
42. Kunkel TA (2009) Evolving views of DNA replication (In)fidelity. *Cold Spring Harbour Symp Quant Biol* 74:91–101
43. Hussain SP, He P, Subleski J, Hofseth LJ, Trivers GE, Mechanic L, Hofseth AB, Bernard M, Schwank J, Nguyen G, Mathe E, Djurickovic D, Haines D, Weiss J, Back T, Gruys E, Laubach VE, Wiltout RH, Harris CC (2008) Nitric oxide is a key component in inflammation-accelerated tumorigenesis. *Cancer Res* 68(17):7130–7136
44. Fearon ER, Vogelstein B (1990) A genetic model for colorectal tumorigenesis. *Cell* 61(5):759–767
45. Blanquet V, Turleau C, Gross-Morand MS, Senamaud-Beaufort C, Doz F, Besmond C (1995) Spectrum of germline mutations in the RB1 gene: a study of 232 patients with hereditary and non hereditary retinoblastoma. *Hum Mol Genet* 4(3):383–388
46. Hussain SP, Harris CC (1998) Molecular epidemiology of human cancer: contribution of mutation spectra studies of tumor suppressor genes. *Cancer Res* 58(18):4023–4037
47. Wilson VL (2001) Detecting rare mutations associated with cancer risk. *Am J Pharmacogenomics* 1(4):283–293
48. Tapp RA, Feng J, Jones JW, Carlson JA, Wilson VL (2007) Single base instability is promoted in vulvar lichen sclerosis. *J Invest Dermatol* 127(11):2563–2576
49. Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, Hainaut P, Olivier M (2007) Impact of mutant p53 functional properties on TP53 mutation patterns and tumor phenotype: lessons from recent developments in the IARC TP53 database. *Hum Mutat* 28(6):622–629
50. Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN (2009) The Human Gene Mutation Database: 2008 update. *Genome Med* 1(1):13
51. Kumar R, Angelini S, Hemminki K (2003) Activating BRAF and N-Ras mutations in sporadic primary melanomas: an inverse association with allelic loss on chromosome 9. *Oncogene* 22(58):9217–9224
52. Aplan PD (2006) Causes of oncogenic chromosomal translocation. *Trends Genet* 22(1):46–55
53. MacCorkle RA, Slattery SD, Nash DR, Brinkley BR (2006) Intracellular protein binding to asbestos induces aneuploidy in human lung fibroblasts. *Cell Motil Cytoskeleton* 63(10):646–657

54. Pavanello S, Pesatori AC, Dioni L, Hoxha M, Bollati V, Siwinska E, Mielzynska D, Bolognesi C, Bertazzi PA, Baccarelli A (2010) Shorter telomere length in peripheral blood lymphocytes of workers exposed to polycyclic aromatic hydrocarbons. *Carcinogenesis* 31(2):216–221
55. Hirao T, Nelson HH, Ashok TD, Wain JC, Mark EJ, Christiani DC, Wiencke JK, Kelsey KT (2001) Tobacco smoke-induced DNA damage and an early age of smoking initiation induce chromosome loss at 3p21 in lung cancer. *Cancer Res* 61(2):612–615
56. Saletta F, Matullo G, Manuguerra M, Arena S, Bardelli A, Vineis P (2007) Exposure to the tobacco smoke constituent 4-aminobiphenyl induces chromosomal instability in human cancer cells. *Cancer Res* 67(15):7088–7094
57. Shinmura K, Iwaizumi M, Igarashi H, Nagura K, Yamada H, Suzuki M, Fukasawa K, Sugimura H (2008) Induction of centrosome amplification and chromosome instability in p53-deficient lung cancer cells exposed to benzo[a]pyrene diol epoxide (B[a]PDE). *J Pathol* 216(3):365–374
58. Groopman JD, Donahue PR, Zhu JQ, Chen JS, Wogan GN (1985) Aflatoxin metabolism in humans: detection of metabolites and nucleic acid adducts in urine by affinity chromatography. *Proc Natl Acad Sci USA* 82(19):6492–6496
59. Groopman JD, Wild CP, Hasler J, Junshi C, Wogan GN, Kensler TW (1993) Molecular epidemiology of aflatoxin exposures: validation of aflatoxin-N7-guanine levels in urine as a biomarker in experimental rat models and humans. *Environ Health Perspect* 99:107–113
60. Essigmann JM, Wood ML (1993) The relationship between the chemical structures and mutagenic specificities of the DNA lesions formed by chemical and physical mutagens. *Toxicol Lett* 67(1–3):29–39
61. Batzer MA, Deininger PL (2002) Alu repeats and human genomic diversity. *Nat Rev Genet* 3(5):370–379
62. Konkel MK, Batzer MA (2010) A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Semin Cancer Biol* 20(4):211–221
63. Callinan PA, Batzer MA (2006) Retrotransposable elements and human disease. *Genome Dyn* 1:104–115
64. Jones PA, Liang G (2009) Rethinking how DNA methylation patterns are maintained. *Nat Rev Genet* 10(11):805–811
65. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature* 462(7271):315–322
66. Schubeler D (2009) Epigenomics: Methylation matters. *Nature* 462(7271):296–297
67. Guil S, Esteller M (2009) DNA methylomes, histone codes and miRNAs: tying it all together. *Int J Biochem Cell Biol* 41(1):87–95
68. Lopez-Serra L, Esteller M (2008) Proteins that bind methylated DNA and human cancer: reading the wrong words. *Br J Cancer* 98(12):1881–1885
69. Esteller M (2008) Epigenetics in cancer. *N Engl J Med* 358(11):1148–1159
70. Sharma S, Kelly TK, Jones PA (2010) Epigenetics in cancer. *Carcinogenesis* 31(1):27–36
71. Wilson VL, Jones PA (1983) Inhibition of DNA methylation by chemical carcinogens in vitro. *Cell* 32(1):239–246
72. Wilson VL, Smith RA, Longoria J, Liotta MA, Harper CM, Harris CC (1987) Chemical carcinogen-induced decreases in genomic 5-methyldeoxycytidine content of normal human bronchial epithelial cells. *Proc Natl Acad Sci USA* 84(10):3298–3301
73. Chanda S, Dasgupta UB, Guhamazumder D, Gupta M, Chaudhuri U, Lahiri S, Das S, Ghosh N, Chatterjee D (2006) DNA hypermethylation of promoter of gene p53 and p16 in arsenic-exposed people with and without malignancy. *Toxicol Sci* 89(2):431–437
74. Bollati V, Baccarelli A, Hou L, Bonzini M, Fustinoni S, Cavallo D, Byun HM, Jiang J, Marinelli B, Pesatori AC, Bertazzi PA, Yang AS (2007) Changes in DNA methylation patterns in subjects exposed to low-dose benzene. *Cancer Res* 67(3):876–880
75. Baccarelli A, Bollati V (2009) Epigenetics and environmental chemicals. *Curr Opin Pediatr* 21(2):243–251
76. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389(6648):251–260
77. Malik HS, Henikoff S (2003) Phylogenomics of the nucleosome. *Nat Struct Biol* 10(11):882–891
78. Allfrey VG, Faulkner R, Mirsky AE (1964) Acetylation and Methylation of Histones and Their Possible Role in the Regulation of Rna Synthesis. *Proc Natl Acad Sci USA* 51:786–794
79. Kondo Y, Shen L, Yan PS, Huang TH, Issa JP (2004) Chromatin immunoprecipitation microarrays for identification of genes silenced by histone H3 lysine 9 methylation. *Proc Natl Acad Sci USA* 101(19):7398–7403
80. McKittrick E, Gafken PR, Ahmad K, Henikoff S (2004) Histone H3.3 is enriched in covalent modifications associated with active chromatin. *Proc Natl Acad Sci USA* 101(6):1525–1530
81. Kondo Y, Shen L, Ahmed S, Bumber Y, Sekido Y, Haddad BR, Issa JP (2008) Downregulation of histone H3 lysine 9 methyltransferase G9a induces centrosome disruption and chromosome instability in cancer cells. *PLoS ONE* 3(4):e2037
82. Yamada T, Mizuno K, Hirota K, Kon N, Wahls WP, Hartsuiker E, Murofushi H, Shibata T, Ohta K (2004) Roles of histone acetylation and chromatin remodeling factor in a meiotic recombination hotspot. *EMBO J* 23(8):1792–1803
83. Toh GW, O'Shaughnessy AM, Jimeno S, Dobbie IM, Grenon M, Maffini S, O'Rourke A, Lowndes NF (2006) Histone H2A phosphorylation and H3 methylation are required for a novel Rad9 DSB repair function following checkpoint activation. *DNA Repair (Amst)* 5(6):693–703
84. Shahbazian MD, Zhang K, Grunstein M (2005) Histone H2B ubiquitylation controls processive methylation but not monomethylation by Dot1 and Set1. *Mol Cell* 19(2):271–277
85. Iniguez-Lluhi JA (2006) For a healthy histone code, a little SUMO in the tail keeps the acetyl away. *ACS Chem Biol* 1(4):204–206
86. Biel M, Wascholowski V, Giannis A (2005) Epigenetics—an epicenter of gene regulation: histones and histone-modifying enzymes. *Angew Chem Int Ed Engl* 44(21):3186–3216

87. Henikoff S (2005) Histone modifications: combinatorial complexity or cumulative simplicity? *Proc Natl Acad Sci USA* 102(15):5308–5309
88. Talasz H, Lindner HH, Sarg B, Helliger W (2005) Histone H4-lysine 20 monomethylation is increased in promoter and coding regions of active genes and correlates with hyperacetylation. *J Biol Chem* 280(46):38814–38822
89. Schneider J, Wood A, Lee JS, Schuster R, Dueker J, Maguire C, Swanson SK, Florens L, Washburn MP, Shilatifard A (2005) Molecular regulation of histone H3 trimethylation by COM-PASS and the regulation of gene expression. *Mol Cell* 19(6):849–856
90. Khalil AM, Driscoll DJ (2006) Histone H3 lysine 4 dimethylation is enriched on the inactive sex chromosomes in male meiosis but absent on the inactive X in female somatic cells. *Cytogenet Genome Res* 112(1–2):11–15
91. Kondo Y, Shen L, Cheng AS, Ahmed S, Bumber Y, Charo C, Yamochi T, Urano T, Furukawa K, Kwabi-Addo B, Gold DL, Sekido Y, Huang TH, Issa JP (2008) Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation. *Nat Genet* 40(6):741–750
92. Taverna SD, Li H, Ruthenburg AJ, Allis CD, Patel DJ (2007) How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nat Struct Mol Biol* 14(11):1025–1040
93. Kurdastani SK, Grunstein M (2003) Histone acetylation and deacetylation in yeast. *Nat Rev Mol Cell Biol* 4(4):276–284
94. Roth SY, Denu JM, Allis CD (2001) Histone acetyltransferases. *Annu Rev Biochem* 70:81–120
95. Brownell JE, Zhou J, Ranalli T, Kobayashi R, Edmondson DG, Roth SY, Allis CD (1996) Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation. *Cell* 84(6):843–851
96. Finnin MS, Donigian JR, Cohen A, Richon VM, Rifkind RA, Marks PA, Breslow R, Pavletich NP (1999) Structures of a histone deacetylase homologue bound to the TSA and SAHA inhibitors. *Nature* 401(6749):188–193
97. Gray SG, Ekstrom TJ (2001) The human histone deacetylase family. *Exp Cell Res* 262(2):75–83
98. Dokmanovic M, Clarke C, Marks PA (2007) Histone deacetylase inhibitors: overview and perspectives. *Mol Cancer Res* 5(10):981–989
99. Verdin E, Dequiedt F, Kasler HG (2003) Class II histone deacetylases: versatile regulators. *Trends Genet* 19(5):286–293
100. Sengupta N, Seto E (2004) Regulation of histone deacetylase activities. *J Cell Biochem* 93(1):57–67
101. Gutierrez RM, Hnilica LS (1967) Tissue specificity of histone phosphorylation. *Science* 157(794):1324–1325
102. Wei Y, Mizzen CA, Cook RG, Gorovsky MA, Allis CD (1998) Phosphorylation of histone H3 at serine 10 is correlated with chromosome condensation during mitosis and meiosis in Tetrahymena. *Proc Natl Acad Sci USA* 95(13):7480–7484
103. Hirota T, Lipp JJ, Toh BH, Peters JM (2005) Histone H3 serine 10 phosphorylation by Aurora B causes HP1 dissociation from heterochromatin. *Nature* 438(7071):1176–1180
104. Goto H, Tomono Y, Ajiro K, Kosako H, Fujita M, Sakurai M, Okawa K, Iwamatsu A, Okigaki T, Takahashi T, Inagaki M (1999) Identification of a novel phosphorylation site on histone H3 coupled with mitotic chromosome condensation. *J Biol Chem* 274(36):25543–25549
105. Preuss U, Landsberg G, Scheidtmann KH (2003) Novel mitosis-specific phosphorylation of histone H3 at Thr11 mediated by Dlk/ZIP kinase. *Nucleic Acids Res* 31(3):878–885
106. Barratt MJ, Hazzalin CA, Cano E, Mahadevan LC (1994) Mitogen-stimulated phosphorylation of histone H3 is targeted to a small hyperacetylation-sensitive fraction. *Proc Natl Acad Sci USA* 91(11):4781–4785
107. Cheung P, Tanner KG, Cheung WL, Sassone-Corsi P, Denu JM, Allis CD (2000) Synergistic coupling of histone H3 phosphorylation and acetylation in response to epidermal growth factor stimulation. *Mol Cell* 5(6):905–915
108. Clayton AL, Rose S, Barratt MJ, Mahadevan LC (2000) Phosphoacetylation of histone H3 on c-fos- and c-jun-associated nucleosomes upon gene activation. *EMBO J* 19(14):3714–3726
109. Panier S, Durocher D (2009) Regulatory ubiquitylation in response to DNA double-strand breaks. *DNA Repair (Amst)* 8(4):436–443
110. Galanty Y, Belotserkovskaya R, Coates J, Polo S, Miller KM, Jackson SP (2009) Mammalian SUMO E3-ligases PIAS1 and PIAS4 promote responses to DNA double-strand breaks. *Nature* 462(7275):935–939
111. Morris JR, Boutell C, Keppler M, Densham R, Weekes D, Alamshah A, Butler L, Galanty Y, Pangon L, Kiuchi T, Ng T, Solomon E (2009) The SUMO modification pathway is involved in the BRCA1 response to genotoxic stress. *Nature* 462(7275):886–890
112. An JY, Kim EA, Jiang Y, Zakrzewska A, Kim DE, Lee MJ, Mook-Jung I, Zhang Y, Kwon YT (2010) UBR2 mediates transcriptional silencing during spermatogenesis via histone ubiquitination. *Proc Natl Acad Sci USA* 107(5):1912–1917
113. Grazini U, Zanardi F, Citterio E, Casola S, Goding CR, McBlane F (2010) The RING domain of RAG1 ubiquitylates histone H3: a novel activity in chromatin-mediated regulation of V(D)J joining. *Mol Cell* 37(2):282–293
114. Lu LY, Wu J, Ye L, Gavrilina GB, Saunders TL, Yu X (2010) RNF8-dependent histone modifications regulate nucleosome removal during spermatogenesis. *Dev Cell* 18(3):371–384
115. Mabruk MJ, Toh LK, Murphy M, Leader M, Kay E, Murphy GM (2009) Investigation of the effect of UV irradiation on DNA damage: comparison between skin cancer patients and normal volunteers. *J Cutan Pathol* 36(7):760–765
116. Wogan GN, Hecht SS, Felton JS, Conney AH, Loeb LA (2004) Environmental and chemical carcinogenesis. *Semin Cancer Biol* 14(6):473–486
117. Lee WR, Perantie DC, Clark KB, Guillot DA, Wilson VL (2001) Effect of mutagen-induced cell lethality on the dose

- response of germline mutations. *Environ Mol Mutagen* 37(4):340–344
118. Ellison KS, Dogliotti E, Connors TD, Basu AK, Essigmann JM (1989) Site-specific mutagenesis by O6-alkylguanines located in the chromosomes of mammalian cells: influence of the mammalian O6-alkylguanine-DNA alkyltransferase. *Proc Natl Acad Sci USA* 86(22):8620–8624
 119. Dosanjh MK, Singer B, Essigmann JM (1991) Comparative mutagenesis of O6-methylguanine and O4-methylthymine in *Escherichia coli*. *Biochemistry* 30(28):7027–7033
 120. Potten CS, Booth C (2002) Keratinocyte stem cells: a commentary. *J Invest Dermatol* 119(4):888–899
 121. Bapat SA (2007) Evolution of cancer stem cells. *Semin Cancer Biol* 17(3):204–213
 122. Orford KW, Scadden DT (2008) Deconstructing stem cell self-renewal: genetic insights into cell-cycle regulation. *Nat Rev Genet* 9(2):115–128
 123. Yahata T, Muguruma Y, Yumino S, Sheng Y, Uno T, Matsuzawa H, Ito M, Kato S, Hotta T, Ando K (2008) Quiescent human hematopoietic stem cells in the bone marrow niches organize the hierarchical structure of hematopoiesis. *Stem Cells* 26(12):3228–3236
 124. Karpowicz P, Morshead C, Kam A, Jervis E, Ramunas J, Cheng V, van der Kooy D (2005) Support for the immortal strand hypothesis: neural stem cells partition DNA asymmetrically in vitro. *J Cell Biol* 170(5):721–732
 125. Lechler T, Fuchs E (2005) Asymmetric cell divisions promote stratification and differentiation of mammalian skin. *Nature* 437(7056):275–280
 126. Smith GH (2005) Label-retaining epithelial cells in mouse mammary gland divide asymmetrically and retain their template DNA strands. *Development* 132(4):681–687
 127. Cairns J (1975) Mutation selection and the natural history of cancer. *Nature* 255(5505):197–200
 128. Potten CS, Grant HK (1998) The relationship between ionizing radiation-induced apoptosis and stem cells in the small and large intestine. *Br J Cancer* 78(8):993–1003
 129. Lee JM, Dedhar S, Kalluri R, Thompson EW (2006) The epithelial-mesenchymal transition: new insights in signaling, development, and disease. *J Cell Biol* 172(7):973–981
 130. Rubin H (2001) The role of selection in progressive neoplastic transformation. *Adv Cancer Res* 83:159–207
 131. Negrini S, Gorgoulis VG, Halazonetis TD (2010) Genomic instability—an evolving hallmark of cancer. *Nat Rev Mol Cell Biol* 11(3):220–228
 132. Beckman RA, Loeb LA (2006) Efficiency of carcinogenesis with and without a mutator mutation. *Proc Natl Acad Sci USA* 103(38):14140–14145
 133. Chu WK, Hickson ID (2009) RecQ helicases: multifunctional genome caretakers. *Nat Rev Cancer* 9(9):644–654
 134. Lengauer C, Kinzler KW, Vogelstein B (1998) Genetic instabilities in human cancers. *Nature* 396(6712):643–649
 135. Ganem NJ, Godinho SA, Pellman D (2009) A mechanism linking extra centrosomes to chromosomal instability. *Nature* 460(7252):278–282
 136. Weaver BA, Silk AD, Montagna C, Verdier-Pinard P, Cleveland DW (2007) Aneuploidy acts both oncogenically and as a tumor suppressor. *Cancer Cell* 11(1):25–36
 137. Thompson SL, Compton DA (2008) Examining the link between chromosomal instability and aneuploidy in human cells. *J Cell Biol* 180(4):665–672
 138. Torres EM, Williams BR, Amon A (2008) Aneuploidy: cells losing their balance. *Genetics* 179(2):737–746
 139. Hussein MR, Wood GS (2002) Microsatellite instability and its relevance to cutaneous tumorigenesis. *J Cutan Pathol* 29(5):257–267
 140. Dietmaier W, Wallinger S, Bocker T, Kullmann F, Fishel R, Ruschoff J (1997) Diagnostic microsatellite instability: definition and correlation with mismatch repair protein expression. *Cancer Res* 57(21):4749–4756
 141. Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, Burt RW, Meltzer SJ, Rodriguez-Bigas MA, Fodde R, Ranzani GN, Srivastava S (1998) A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res* 58(22):5248–5257
 142. Hoeijmakers JH (2001) Genome maintenance mechanisms for preventing cancer. *Nature* 411(6835):366–374
 143. Jascur T, Boland CR (2006) Structure and function of the components of the human DNA mismatch repair system. *Int J Cancer* 119(9):2030–2035
 144. Thibodeau SN, French AJ, Roche PC, Cunningham JM, Tester DJ, Lindor NM, Moslein G, Baker SM, Liskay RM, Burgart LJ, Honchel R, Halling KC (1996) Altered expression of hMSH2 and hMLH1 in tumors with microsatellite instability and genetic alterations in mismatch repair genes. *Cancer Res* 56(21):4836–4840
 145. Hussein MR, Wood GS (2002) Microsatellite instability in human melanocytic skin tumors: an incidental finding or a pathogenetic mechanism? *J Cutan Pathol* 29(1):1–4
 146. Fishel R, Lescoe MK, Rao MR, Copeland NG, Jenkins NA, Garber J, Kane M, Kolodner R (1993) The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* 75(5):1027–1038
 147. Leach FS, Nicolaides NC, Papadopoulos N, Liu B, Jen J, Parsons R, Peltomaki P, Sistonen P, Aaltonen LA, Nystrom-Lahti M et al (1993) Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* 75(6):1215–1225
 148. Garcon G, Dagher Z, Zerimech F, Ledoux F, Courcot D, Aboukais A, Puskaric E, Shirali P (2006) Dunkerque City air pollution particulate matter-induced cytotoxicity, oxidative stress and inflammation in human epithelial lung cells (L132) in culture. *Toxicol In Vitro* 20(4):519–528
 149. Saint-Georges F, Garcon G, Escande F, Abbas I, Verdin A, Gosset P, Mulliez P, Shirali P (2009) Role of air pollution Particulate Matter (PM_{2.5}) in the occurrence of loss of heterozygosity in multiple critical regions of 3p chromosome in human epithelial lung cells (L132). *Toxicol Lett* 187(3):172–179

150. Hirose T, Kondo K, Takahashi Y, Ishikura H, Fujino H, Tsuyuguchi M, Hashimoto M, Yokose T, Mukai K, Kodama T, Monden Y (2002) Frequent microsatellite instability in lung cancer from chromate-exposed workers. *Mol Carcinog* 33(3):172–180
151. Zienolddiny S, Aguelon AM, Mironov N, Mathew B, Thomas G, Sankaranarayanan R, Yamasaki H (2004) Genomic instability in oral squamous cell carcinoma: relationship to betel-quid chewing. *Oral Oncol* 40(3):298–303
152. Parsons BL, Marchant-Miros KE, Delongchamp RR, Verkler TL, Patterson TA, McKinzie PB, Kim LT (2010) ACB-PCR quantification of K-RAS codon 12 GAT and GTT mutant fraction in colon tumor and non-tumor tissue. *Cancer Invest* 28(4):364–375
153. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa JP (1999) CpG island methylator phenotype in colorectal cancer. *Proc Natl Acad Sci USA* 96(15):8681–8686
154. Teodoridis JM, Hardie C, Brown R (2008) CpG island methylator phenotype (CIMP) in cancer: causes and implications. *Cancer Lett* 268(2):177–186
155. Tanemura A, Terando AM, Sim MS, van Hoesel AQ, de Maat MF, Morton DL, Hoon DS (2009) CpG island methylator phenotype predicts progression of malignant melanoma. *Clin Cancer Res* 15(5):1801–1807
156. Yagi K, Akagi K, Hayashi H, Nagae G, Tsuji S, Isagawa T, Midorikawa Y, Nishimura Y, Sakamoto H, Seto Y, Aburatani H, Kaneda A (2010) Three DNA methylation epigenotypes in human colorectal cancer. *Clin Cancer Res* 16(1):21–33
157. Gaudet F, Hodgson JG, Eden A, Jackson-Grusby L, Dausman J, Gray JW, Leonhardt H, Jaenisch R (2003) Induction of tumors in mice by genomic hypomethylation. *Science* 300(5618):489–492
158. Gama-Sosa MA, Slagel VA, Trewyn RW, Oxenhandler R, Kuo KC, Gehrke CW, Ehrlich M (1983) The 5-methylcytosine content of DNA from human tumors. *Nucleic Acids Res* 11(19):6883–6894
159. Goelz SE, Vogelstein B, Hamilton SR, Feinberg AP (1985) Hypomethylation of DNA from benign and malignant human colon neoplasms. *Science* 228(4696):187–190
160. Kim YI, Giuliano A, Hatch KD, Schneider A, Nour MA, Dallal GE, Selhub J, Mason JB (1994) Global DNA hypomethylation increases progressively in cervical dysplasia and carcinoma. *Cancer* 74(3):893–899
161. Shahrzad S, Bertrand K, Minhas K, Coomber BL (2007) Induction of DNA hypomethylation by tumor hypoxia. *Epigenetics* 2(2):119–125
162. Avila MA, Carretero MV, Rodriguez EN, Mato JM (1998) Regulation by hypoxia of methionine adenosyltransferase activity and gene expression in rat hepatocytes. *Gastroenterology* 114(2):364–371
163. Yanagawa N, Tamura G, Honda T, Endoh M, Nishizuka S, Motoyama T (2004) Demethylation of the synuclein gamma gene CpG island in primary gastric cancers and gastric cancer cell lines. *Clin Cancer Res* 10(7):2447–2451
164. Laner T, Schulz WA, Engers R, Muller M, Florl AR (2005) Hypomethylation of the XIST gene promoter in prostate cancer. *Oncol Res* 15(5):257–264
165. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suner D, Cigudosa JC, Urioste M, Benitez J, Boix-Chornet M, Sanchez-Aguilera A, Ling C, Carlsson E, Poulsen P, Vaag A, Stephan Z, Spector TD, Wu YZ, Plass C, Esteller M (2005) Epigenetic differences arise during the lifetime of monozygotic twins. *Proc Natl Acad Sci USA* 102(30):10604–10609
166. Kampranis SC, Tschlis PN (2009) Histone demethylases and cancer. *Adv Cancer Res* 102:103–169
167. Kristensen LS, Nielsen HM, Hansen LL (2009) Epigenetics and cancer treatment. *Eur J Pharmacol* 625(1–3):131–142
168. Kinzler KW, Vogelstein B (1997) Cancer-susceptibility genes. Gatekeepers and caretakers. *Nature* 386(6627):761–763
169. Bardelli A, Cahill DP, Lederer G, Speicher MR, Kinzler KW, Vogelstein B, Lengauer C (2001) Carcinogen-specific induction of genetic instability. *Proc Natl Acad Sci USA* 98(10):5770–5775
170. Palmero EI, Achatz MI, Ashton-Prolla P, Olivier M, Hainaut P (2010) Tumor protein 53 mutations and inherited cancer: beyond Li-Fraumeni syndrome. *Curr Opin Oncol* 22(1):64–69
171. Amberger J, Bocchini CA, Scott AF, Hamosh A (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res* 37(Database issue): D793–D796
172. Cazier JB, Tomlinson I (2010) General lessons from large-scale studies to identify human cancer predisposition genes. *J Pathol* 220(2):255–262
173. Gresner P, Gromadzinska J, Wasowicz W (2007) Polymorphism of selected enzymes involved in detoxification and biotransformation in relation to lung cancer. *Lung Cancer* 57(1):1–25
174. Cote ML, Wenzlaff AS, Bock CH, Land SJ, Santer SK, Schwartz DR, Schwartz AG (2007) Combinations of cytochrome P-450 genotypes and risk of early-onset lung cancer in Caucasians and African Americans: a population-based study. *Lung Cancer* 55(3):255–262
175. Schwartz AG, Prysak GM, Bock CH, Cote ML (2007) The molecular epidemiology of lung cancer. *Carcinogenesis* 28(3):507–518
176. Cokkinides V, Bandi P, McMahon C, Jemal A, Glynn T, Ward E (2009) Tobacco control in the United States—recent progress and opportunities. *CA Cancer J Clin* 59(6):352–365
177. Lubin JH, Alavanja MC, Caporaso N, Brown LM, Brownson RC, Field RW, Garcia-Closas M, Hartge P, Hauptmann M, Hayes RB, Kleinerman R, Kogevinas M, Krewski D, Langholz B, Letourneau EG, Lynch CF, Malats N, Sandler DP, Schaffrath-Rosario A, Schoenberg JB, Silverman DT, Wang Z, Wichmann HE, Wilcox HB, Zielinski JM (2007) Cigarette smoking and cancer risk: modeling total exposure and intensity. *Am J Epidemiol* 166(4):479–489
178. Robles AI, Linke SP, Harris CC (2002) The p53 network in lung carcinogenesis. *Oncogene* 21(45):6898–6907
179. Olivier M, Hollstein M, Hainaut P (2010) TP53 Mutations in Human Cancers: Origins, Consequences, and Clinical Use. *Cold Spring Harb Perspect Biol* 2(1):a001008
180. Denissenko MF, Pao A, Tang M, Pfeifer GP (1996) Preferential formation of benzo[a]pyrene adducts at lung cancer mutational hotspots in P53. *Science* 274(5286):430–432

181. Greenblatt MS, Bennett WP, Hollstein M, Harris CC (1994) Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis. *Cancer Res* 54(18):4855–4878
182. Bennett WP, Hussain SP, Vahakangas KH, Khan MA, Shields PG, Harris CC (1999) Molecular epidemiology of human cancer risk: gene-environment interactions and p53 mutation spectrum in human lung cancer. *J Pathol* 187(1):8–18
183. Hsu IC, Metcalf RA, Sun T, Welsh JA, Wang NJ, Harris CC (1991) Mutational hotspot in the p53 gene in human hepatocellular carcinomas. *Nature* 350(6317):427–428
184. Soini Y, Chia SC, Bennett WP, Groopman JD, Wang JS, DeBenedetti VM, Cawley H, Welsh JA, Hansen C, Bergasa NV, Jones EA, DiBisceglie AM, Trivers GE, Sandoval CA, Calderon IE, Munoz Espinosa LE, Harris CC (1996) An aflatoxin-associated mutational hotspot at codon 249 in the p53 tumor suppressor gene occurs in hepatocellular carcinomas from Mexico. *Carcinogenesis* 17(5):1007–1012
185. Gouas D, Shi H, Hainaut P (2009) The aflatoxin-induced TP53 mutation at codon 249 (R249S): biomarker of exposure, early detection and target for therapy. *Cancer Lett* 286(1):29–37
186. Brash DE, Rudolph JA, Simon JA, Lin A, McKenna GJ, Baden HP, Halperin AJ, Ponten J (1991) A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. *Proc Natl Acad Sci USA* 88(22):10124–10128
187. Armstrong BK, Kricker A (2001) The epidemiology of UV induced skin cancer. *J Photochem Photobiol B* 63(1–3):8–18
188. Runger TM (2007) How different wavelengths of the ultraviolet spectrum contribute to skin carcinogenesis: the role of cellular damage responses. *J Invest Dermatol* 127(9):2103–2105
189. Fang JL, Vaca CE (1997) Detection of DNA adducts of acetaldehyde in peripheral white blood cells of alcohol abusers. *Carcinogenesis* 18(4):627–632
190. Vaca CE, Nilsson JA, Fang JL, Grafstrom RC (1998) Formation of DNA adducts in human buccal epithelial cells exposed to acetaldehyde and methylglyoxal in vitro. *Chem Biol Interact* 108(3):197–208
191. Clemens DL, Forman A, Jerrells TR, Sorrell MF, Tuma DJ (2002) Relationship between acetaldehyde levels and cell survival in ethanol-metabolizing hepatoma cells. *Hepatology* 35(5):1196–1204
192. Molina PE, Hoek JB, Nelson S, Guidot DM, Lang CH, Wands JR, Crawford JM (2003) Mechanisms of alcohol-induced tissue injury. *Alcohol Clin Exp Res* 27(3):563–575
193. el Ghissassi F, Barbin A, Nair J, Bartsch H (1995) Formation of 1,N6-ethenoadenine and 3,N4-ethenocytosine by lipid peroxidation products and nucleic acid bases. *Chem Res Toxicol* 8(2):278–283
194. Giovannucci E (2004) Alcohol, one-carbon metabolism, and colorectal cancer: recent insights from molecular studies. *J Nutr* 134(9):2475S–2481S
195. Giovannucci E (2002) Epidemiologic studies of folate and colorectal neoplasia: a review. *J Nutr* 132(8 Suppl):2350S–2355S
196. Leiter U, Garbe C (2008) Epidemiology of melanoma and nonmelanoma skin cancer—the role of sunlight. *Adv Exp Med Biol* 624:89–103
197. Ridley AJ, Whiteside JR, McMillan TJ, Allinson SL (2009) Cellular and sub-cellular responses to UVA in relation to carcinogenesis. *Int J Radiat Biol* 85(3):177–195
198. Matsunaga T, Hieda K, Nikaido O (1991) Wavelength dependent formation of thymine dimers and (6–4) photoproducts in DNA by monochromatic ultraviolet light ranging from 150 to 365 nm. *Photochem Photobiol* 54(3):403–410
199. Lacour JP (2002) Carcinogenesis of basal cell carcinomas: genetics and molecular mechanisms. *Br J Dermatol* 146(Suppl 61):17–19
200. Agar NS, Halliday GM, Barnetson RS, Ananthaswamy HN, Wheeler M, Jones AM (2004) The basal layer in human squamous tumors harbors more UVA than UVB fingerprint mutations: a role for UVA in human skin carcinogenesis. *Proc Natl Acad Sci USA* 101(14):4954–4959
201. Kuhn C, Hurwitz SA, Kumar MG, Cotton J, Spandau DF (1999) Activation of the insulin-like growth factor-1 receptor promotes the survival of human keratinocytes following ultraviolet B irradiation. *Int J Cancer* 80(3):431–438
202. Lewis DA, Spandau DF (2007) UVB activation of NF-kappaB in normal human keratinocytes occurs via a unique mechanism. *Arch Dermatol Res* 299(2):93–101
203. Lewis DA, Yi Q, Travers JB, Spandau DF (2008) UVB-induced senescence in human keratinocytes requires a functional insulin-like growth factor-1 receptor and p53. *Mol Biol Cell* 19(4):1346–1353
204. Tavakkol A, Elder JT, Griffiths CE, Cooper KD, Talwar H, Fisher GJ, Keane KM, Foltin SK, Voorhees JJ (1992) Expression of growth hormone receptor, insulin-like growth factor 1 (IGF-1) and IGF-1 receptor mRNA and proteins in human skin. *J Invest Dermatol* 99(3):343–349
205. Ferber A, Chang C, Sell C, Ptasznik A, Cristofalo VJ, Hubbard K, Ozer HL, Adamo M, Roberts CT Jr, LeRoith D et al (1993) Failure of senescent human fibroblasts to express the insulin-like growth factor-1 gene. *J Biol Chem* 268:17883–17888
206. Agents Classified by the IARC Monographs 1–100, Updated 30 August 2010; IARC Monographs on the Evaluation of Carcinogenic Risks to Humans. <http://monographs.iarc.fr/ENG/Classification/index.php>. Accessed November 2010
207. Hecht SS, Carmella SG, Murphy SE, Akerkar S, Brunnemann KD, Hoffmann D (1993) A tobacco-specific lung carcinogen in the urine of men exposed to cigarette smoke. *N Engl J Med* 329(21):1543–1546
208. Smoking and Tobacco Control Monograph 10, Health Effects of Exposure to Environmental Tobacco Smoke (August 1999), National Cancer Institute, National Institutes of Health, Bethesda, MD
209. Alexandrov K, Rojas M, Kadlubar FF, Lang NP, Bartsch H (1996) Evidence of anti-benzo[a]pyrene diolepoxide-DNA adduct formation in human colon mucosa. *Carcinogenesis* 17(9):2081–2083

210. Hecht SS, Hoffmann D (1988) Tobacco-specific nitrosamines, an important group of carcinogens in tobacco and tobacco smoke. *Carcinogenesis* 9(6):875–884
211. Pleasance ED, Stephens PJ, O'Meara S, McBride DJ, Meynert A, Jones D, Lin ML, Beare D, Lau KW, Greenman C, Varela I, Nik-Zainal S, Davies HR, Ordóñez GR, Mudie LJ, Latimer C, Edkins S, Stebbings L, Chen L, Jia M, Leroy C, Marshall J, Menzies A, Butler A, Teague JW, Mangion J, Sun YA, McLaughlin SF, Peckham HE, Tsung EF, Costa GL, Lee CC, Minna JD, Gazdar A, Birney E, Rhodes MD, McKernan KJ, Stratton MR, Futreal PA, Campbell PJ (2010) A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 463(7278):184–190
212. McCullough ML, Giovannucci EL (2004) Diet and cancer prevention. *Oncogene* 23(38):6349–6364
213. Kensler TW, Groopman JD (2009) Is it time to advance the chemoprevention of environmental carcinogenesis with microdosing trials? *Cancer Prev Res (Phila Pa)* 2(12):1003–1007
214. Ju J, Picinich SC, Yang Z, Zhao Y, Suh N, Kong AN, Yang CS (2010) Cancer-preventive activities of tocopherols and tocotrienols. *Carcinogenesis* 31(4):533–542
215. Fowke JH (2007) Head and neck cancer: a case for inhibition by isothiocyanates and indoles from cruciferous vegetables. *Eur J Cancer Prev* 16(4):348–356
216. Singh N, Manshian B, Jenkins GJ, Griffiths SM, Williams PM, Maffei TG, Wright CJ, Doak SH (2009) NanoGenotoxicology: the DNA damaging potential of engineered nanomaterials. *Biomaterials* 30(23–24):3891–3914

Environmental Toxicology: Children at Risk

MARIA J. CARROQUINO¹, M. POSADA¹, P. J. LANDRIGAN²

¹Instituto de Salud Carlos III, Madrid, Spain

²Department of Preventive Medicine, Mount Sinai School of Medicine, New York, NY, USA

Article Outline

Glossary

Definition of the Subject

Introduction: Brief History of the Issue of Children's Vulnerability to Environmental Hazards

Toxicological Basis of Children's Vulnerability to

Environmental Hazards: Susceptibility and

Exposure Factors Affecting Children's Vulnerability

Major Groups of Pollutants to Which Children are Exposed

Environmental Threats to Children on Specific Organ Systems

Future Directions

Bibliography

Glossary

Adult The time of life usually starting at 18 years (some systems such as skeleton and brain may continue to develop).

Adverse effect A treatment-related alteration from baseline that diminishes an organism's ability to survive, reproduce, or adapt to the environment.

Carcinogen Any substance that can cause cancer.

Critical period A specific phase during which a developing system is particularly susceptible.

Developmental disorders/effects Adverse effects such as altered growth, structural abnormality, functional deficiency, or death observed in a developing organism.

Dose (exposure)–response relationship Characterization of the relationship between administered dose or exposure and the biological change in organisms.

Embryonic period The period from fertilization to the end of major organogenesis.

Environmental exposures and environmental hazards For this entry, these terms refer to specific environmental chemicals and environmental pollutants.

Exposure Contact with a chemical by swallowing, by breathing, or by direct contact (such as through the skin or eyes). Exposure may be short term (acute) or long term (chronic).

Fetus The period from 8 weeks of pregnancy to birth.

Gestation Length of time between conception and birth.

Infant The period from 28 days of age to 1 year.

Lowest observed adverse effect level (LOAEL) The lowest concentration of a chemical in a study, or group of studies, that produces statistically or biologically significant increase in the frequency or severity of adverse effects between the exposed population and an appropriate control.

Mechanism of action The detailed molecular knowledge of the key events leading to an adverse effect in an organism.

Neonate The period from birth to 28 days of age.

Perinatal stage The period of 29 weeks of pregnancy to 7 days after birth.

Pregnancy The condition of having an implanted embryo or fetus in the body, after fusion of an ovum and spermatozoon.

Preterm birth A birth occurring at 24–37 weeks of pregnancy.

Risk assessment An empirically based paradigm that estimates the risk of adverse effect(s) from exposure of an individual or population to a chemical, physical, or biological agent. It includes the components of hazard identification, assessment of dose–response relationships, exposure assessment, and risk characterization.

Route of exposure Exposure route refers to the different ways a substance may enter the body. The route may be dermal, ingestion, or inhalation.

Sexual maturation Achievement of full development of the reproductive system and sexual function.

Susceptibility An individual's intrinsic or acquired traits that modify the risk of illness (e.g., high susceptibility to cancer).

Toxicokinetics The process of the uptake of potentially toxic substances by the body, the biotransformation they undergo, the distribution of the substances and their metabolites in the tissues, and the elimination of the substances and their metabolites from the body. Both the amounts and the concentrations of the substances and their metabolites are studied. (Pharmacokinetics is the term used to study pharmaceutical substances.)

Vulnerability A matrix of physical, chemical, biological, social, and cultural factors that result in certain communities and subpopulations being more susceptible to environmental factors because of greater exposure to such factors or a compromised ability to cope with and/or recover from such exposure.

Definition of the Subject

Children today live in a world that is vastly different from a few generations ago. While industrialization has maximized (for many) children's opportunities to survive, develop and enjoy high levels of health, education, recreation, and fulfillment, it has also added significant challenges to their development.

In the countries in the Organization for Economic Co-operation and Development (OECD), an infant born today has about 20 years longer life expectancy than one born at the beginning of the twentieth century [1]. Infant mortality has gone down by over 90%. Despite AIDS, SARS, West Nile virus, and the constant threat of other emerging infections, the ancient epidemics of smallpox, yellow fever, cholera, bubonic plague, polio, and measles are no longer the dominant causes of disease and death. However, an increase in the incidence (new cases) of many chronic diseases in children has been observed. These include asthma, cancer (which is the second leading cause of death in children after injuries), birth defects, developmental disabilities, and autism. For some of these conditions, an environmental origin has been established, and for others it is hypothesized.

Environmental threats to children's health range from asthma-inducing air pollution and lead-based paint in older homes, to treatment-resistant microbes in drinking water and persistent industrial chemicals that may cause cancer or result in reproductive or developmental changes.

Patterns of illness among children in the industrially developed nations have changed substantially. Infant mortality has declined. Life expectancy has increased. However, an increase in the incidence (new cases) of some diseases, such as asthma, birth defects, neurodevelopmental disorders and certain types of childhood and adolescence cancers, and obesity has been observed. Evidence is accumulating that toxic chemicals are responsible for at least some of these changing patterns of disease.

Children are uniquely susceptible to chemicals. This great vulnerability reflects the juxtaposition of two phenomena early in life: first, that infants and children have disproportionately greater exposures than adults to environmental chemicals, and second, that children are exquisitely sensitive to these exposures, because they are poorly equipped to metabolize many toxic compounds and because they are progressing through the complex, delicate, and easily disrupted stages of early development. The protection of children from environmental health hazards requires the consideration of their exposure patterns and susceptibility factors when conducting risk assessments, development of child protective legislation, and wider

application of the Precautionary Principle in the face of early warning of danger.

Introduction: Brief History of the Issue of Children's Vulnerability to Environmental Hazards

Children's risks from environmental health threats have received considerable political attention during the last two decades. In 1989, the United Nations Convention on the Rights of the Child laid down basic standards for the protection of children and proclaimed that they are entitled to special care and assistance. A year later, the World Summit for Children (WSC) adopted a Declaration on the Survival, Protection and Development for Children, in which the signatories agreed to join efforts on taking measures to protect the environment, so that all children can enjoy a safer and healthier future.

Agenda 21, adopted in 1992 at the United Nations Conference on Environment and Development ("the Earth Summit") gives attention to the protection of children from the effects of a deteriorating environment in several chapters. Chapter 6 of Agenda 21 "Protecting and Promoting Human Health" emphasizes the need to pay special attention to protecting vulnerable groups, particularly infants, young people, women, indigenous people, and the poor. Agenda 21 urges governments to develop programs to protect children from the effects of environmental and occupational toxic compounds.

In the USA, the National Research Council report [2] "Pesticides in the Diet of Infants and Children" was critical in raising awareness about the importance in risk assessment of children's environmental health. This report elevated concern on a broad national level about children's special vulnerabilities to environmental agents. It made clear that protection of the health of vulnerable populations would require a new approach to risk assessment. The NRC report recommended an approach to risk assessment that moved beyond consideration of average exposures based primarily on adult characteristics to one that accounted for the heterogeneity of exposures and for potential differential sensitivities of various life stages, particularly during prenatal development, infancy, and childhood.

Responding to recommendations in the NRC report [2], the US Government took decisive steps to

attend to the growing concern on children susceptibility to environmental toxicants. In 1995, the EPA issued a National Policy to consistently and explicitly take into account health risks to children and infants from environmental hazards when conducting assessments of environmental risks. In 1997, the Clinton Administration issued an Executive Order, Executive Order 13045: Protection of Children from Environmental Health Risks and Safety Risks. The Executive Order requires all federal agencies to address health and safety risks to children, to coordinate research priorities on children's health, and to ensure that their standards take into account special risks to children. To implement the order, the US EPA established the Office of Children's Health Protection (OCHP) (renamed the Office of Children's Health Protection and Environmental Education – OCHPEE – in 2005), whose job is to work with Program and Regional Offices within the US EPA to promote a safe and healthy environment for children by ensuring that all regulations, standards, policies, and risk assessments take into account risks to children. Legislation, such as the Food Quality Protection Act and the Safe Drinking Water Act amendments, has made coverage of children's health issues more explicit, and research on children's health issues is continually expanding. As a result of the emphasis on children's risk, the US EPA Office of Research and Development (ORD) developed a strategy for research on environmental risks to children. The goal of this research agenda is to discover the environmental causes of disease in children and then to convert these research findings into blueprints for disease prevention and health promotion.

The 1997 Declaration of the Environmental Leaders of the Eight on Children's Environmental Health intensified their commitment to protecting children's health from environmental hazards. The Environment Ministers of the G8 countries acknowledged the special vulnerabilities of children and committed their countries to taking action on several specific environmental health issues such as lead, microbiologically safe drinking water, endocrine disrupting chemicals, environmental tobacco smoke (ETS), and air quality. They called on financial institutions, the World Health Organization (WHO), the United Nations Environment Program (UNEP), and other international bodies to continue ongoing

activities and to pay further attention to children's environmental health, in particular the economic and social dimensions of children's health. In addition, they committed their countries to fulfilling and to promoting the Organization for Economic Cooperation and Development (OECD) Declaration on Risk Reduction for Lead. The underlying rationale for each of these actions is that disease of environmental origin in children can be prevented by controlling harmful exposures in the environment.

International organizations such as the World Health Organization (WHO), the United Nations Environment Program (UNEP), and the European Union (EU) responded to this call and also developed initiatives on Children, Environment and Health. Children's Health and Environment has been the central focus of Europe wide Ministerial Conferences organized by WHO, and other EU initiatives, such as SCALE (Science, Children, Awareness, Legal Instrument, Evaluation). Thus, the protection of children from environmental hazards has been a subject of intense political and scientific attention and effort during the last two decades. Since the initial reports on this issue, research has intensified, and scientific literature abounds with new findings on children's differential vulnerability to environmental health threats that have provided the basis to maintaining the political attention on children's health.

Toxicological Basis of Children's Vulnerability to Environmental Hazards: Susceptibility and Exposure Factors Affecting Children's Vulnerability

Over the past 2 decades, a number of scientific reports have integrated the knowledge available on the susceptibility of children to environmental toxicants, have highlighted knowledge gaps, and have pointed to future directions of research to fill these gaps. This has expanded and deepened the knowledge of the differential susceptibility of children to environmental toxicants, and at the same time has allowed the firm establishment of certain fundamentals concerning children's susceptibility to environmental toxicants.

The International Life Science Institute (ILSI) and the EPA, recognizing the need for examining the scientific evidence on the broad question of the differential

susceptibility of children to environmental hazards, organized in 1990 a conference titled: "Similarities and Differences Between Children and Adults: Implications for Risk Assessment." The results from this conference were summarized in the publication "Similarities and Differences Between Children and Adults" [3] and was incorporated into ongoing work of the National Research Committee on Pesticides in the Diets of Infants and Children. The National Academy of Science (NAS) published the NRC report "Pesticides in the Diets of Infants and Children" [2], which was critical in raising awareness of the importance of considering the vulnerable life stages of children when conducting risk assessment of exposures to children. In 2001, the International Life Sciences Institute convened a number of scientific experts to develop a conceptual framework for conducting health risk assessments of children exposures, which takes into consideration their unique characteristics and special vulnerabilities [4].

The ILSI report highlighted the fact that "children are not little adults," but a unique subpopulation that needs to be considered in risk assessment due to differential exposure patterns, immaturity in physiological development, or differential toxicant metabolism.

In 2006, the WHO published the Principles for Evaluating the Health Risks in Children Associated with Exposure to Chemicals in the Environmental Health Criteria 237 [5]. This publication constitutes the most recent monographic work of the subject of children susceptibility to environmental toxicants, and provides comprehensive information on children's developmental stages and the critical windows of susceptibility that appear through the course of their development.

In summary, it is clear from these three major scientific reports – NAS, ILSI, and WHO – that there is broad scientific consensus that risks to children from environmental health threats differ qualitatively or quantitatively from those of adults for a variety of reasons:

- Differential exposure patterns: Compared to adults, children have heavier exposures in relation to body weight because they drink more water, eat more food, and have higher breathing rates per unit of body weight than adults do. As a consequence, children have substantially heavier

exposures than adults to any toxicants that are present in water, food, or air.

- Children's ability to metabolize, detoxify, and excrete chemicals is different from that of adults. During the first months after birth, their metabolic pathways are immature. In some cases, children may actually have a higher metabolic capacity for some toxicants than adults. Commonly, however, they are less able to deal with toxic chemicals and thus are more vulnerable to them.
- Children undergo rapid growth and development, and their development phases are perfectly scheduled to achieve complete functional development. If a development phase is disturbed at a given time, its time to take place may be lost definitively. Thus, interferences with certain phases of development may have irreversible effects. If cells in a child's developing brain are destroyed by chemicals such as lead, mercury, or solvents, or if false signals are sent to the developing reproductive organs by endocrine disruptors, there is a high risk that the resulting dysfunction will be permanent and irreversible.
- Children's exposures are affected qualitatively and quantitatively by their behavior and the unique microenvironment in which they spend their time. Their hand-to-mouth behavior brings contaminated items or soil to their mouth, and living and playing closer to the ground also exposes them to pollutants in a different pattern than adults, both in quantitative and qualitative terms.
- Because children generally have more years of life ahead than adults, they have more time to develop chronic diseases triggered by early exposures. Many diseases, such as cancer and neurodegenerative diseases, are thought to arise through a series of stages that require years or even decades from initiation to actual manifestation of disease. Carcinogenic and toxic exposures, sustained early in life, including prenatal exposures, would then be more likely to lead to disease than similar exposures encountered later.

The purpose of this chapter has been to select, integrate, and summarize the most relevant and illustrative information that explains and highlights the potential susceptibility of children to environmental exposures. The goal of the chapter is to promote

understanding of the issue of differential susceptibility and exposures of children to environmental threats among scientists from adjacent fields, and to encourage scientists to consider the issue of children's differential susceptibility and exposures in their research. Recognition of children's exquisite vulnerability to toxic exposures in the environment is critical to child-protective risk assessment and to disease prevention.

For a more in-depth study of any of the particular health issues or particular toxicants, extensive and rigorous reviews that are cited through this document can be consulted. The following sections will elaborate on the general toxicological aspects of children's susceptibility.

Exposure

There are several factors that influence the differential exposures of children to environmental toxicants. These are: (1) their higher ingestion, drinking, and breathing rates per unit of body weight, (2) their unique behaviors, (3) where they spend their time, and (4) their unique microenvironments.

Ingestion, Breathing, and Drinking Rates In toxicology, exposure is defined as the contact that occurs between a receptor (a living organism) and an environmental agent. The most commonly considered exposure pathways are inhalation, ingestion, and dermal absorption. For infants in the womb, transplacental transfer of toxic chemicals is another unique route of exposure, and for nursing infants, the proportion of the toxicant that enters the body through each of these pathways will depend on both external, environmental factors, and on biological factors that will determine the rate of uptake of the chemical by the organism. Exposure can also be referred to as the concentration that a target organ receives, once the chemical has been absorbed.

The environmental concentration and duration of exposure will determine the extent of exposure. In addition, other environmental factors such as for example, temperature, humidity, and pH, may alter pollutant concentrations. For example, the concentration of some volatile chemicals may vary under different temperature conditions. Sometimes, chemical reactions between pollutants in the environment may

change qualitatively the nature of exposure, such as in the case of ozone formation in the air from nitrogen oxides (NO_x), carbon monoxide (CO), and volatile organic compounds (VOCs) in the presence of sunlight.

The patterns of exposure to ingested and inhaled toxicants are different between children and adults. Young children drink more water on a body weight basis than adults do (seven times as much water per kilogram of body weight). For the first 6 months, for example, children consume on average 88 mL/kg/day of tap water directly or indirectly (water added in the preparation of formula or fruit juice) compared to adults who drink 17 mL/kg/day [6]. As a result of this greater consumption of water, children may be disproportionately exposed to chemicals found in drinking water, including the water used to make infant formula [2].

Infants and young children have a higher resting metabolic rate and rate of oxygen consumption per unit body weight than adults. The oxygen consumption of a resting infant aged between 1 week and 1 year is 7 mL/kg body weight per minute, compared to that of an adult in the same condition, which is 3–5 L/kg/min [5]. Thus, on a body weight basis, the volume of air passing through the lungs of a resting infant is twice that of an adult, and therefore twice as much of any chemical in the atmosphere would pass through the lungs of an infant. An additional consideration is the smaller lung surface area per kilogram in the early stages of development. Thus, the higher amount of inspired air will affect a relatively smaller area of lung tissue. In addition, children have narrower airways than those of adults. Thus, irritation caused by air pollution that would produce only a slight response in an adult, can result in potentially significant obstruction in the airways of a young child [6]. Furthermore, the fact that children spend more time engaged in vigorous activities than adults exacerbates the differential effects. Children also ingest more food per unit of body weight than do adults. A 1–5 year old child eats three to four times more food per kilogram than the average adult, resulting in larger amounts of chemicals and infectious agents per unit of body mass [2, 7]. A child's diet is very different from an adult's. The diet of children contains more milk products and more fruits and vegetables per unit of body weight

than adults. During the first year of life, human milk or cow milk-based products constitute most of their energy and nutrient source. The NRC report [2] reported that over the first year, cow milk products comprise 36% and 58% of the diets of nursing and non-nursing infants, respectively, compared to adults, where milk and milk products constitute only about 29% of their diet. Environmental pollutants, such as PCBs and dioxins, accumulate in fat and have been found in breast milk, although it is broadly accepted that the benefits of breast-feeding still outweigh the risks of exposure [5].

After the first few months of life, fruits and fruit juices constitute a large proportion of infants' diet. For example, the average consumption of apples for children between birth and 5 months of age is almost ten times that of an adult older than 20 years old [6]. Fruits can be contaminated with pesticides and other toxicants, causing children to be exposed to these chemicals.

Behavioral Patterns Influencing Exposure in Children

Two characteristic behaviors of children have been considered and studied in relation to children's exposure to chemicals: mouthing behavior and pica behavior. Infants and toddlers pass through a developmental phase, characterized by intense oral exploratory behavior, when they indiscriminately bring their hands and objects to their mouths. Mouthing behavior can result in oral exposures to chemicals that may be part of the surfaces of objects, or adhered to them in dust particles.

Pica is a behavior of some children that has been defined as the craving or ingestion of nonfood items. The cravings found in patients with pica have been associated with a nutritional deficiency state, such as iron-deficient anemia; with pregnancy or with mental illness. Pica tends to disappear as children grow older, except for mentally retarded children [8]. Pica behavior can give rise to the ingestion of soil, paint, and other possibly contaminated substances. About 95% of children ingest 0.2 g of soil per day or less, but studies have shown that children with pica behavior ingest up to 60 g of soil per day [8]. Children living in neighborhoods built on soil contaminated with heavy metals (such as lead), aromatic hydrocarbons, and other compounds were found to be at increased risk for elevated blood levels [8].

Children may also have higher exposures because of their higher levels of physical activity, at least during certain stages of their development. When they are outdoors, they are more prone to run, jump, or play vigorously than adults do. Their breathing rates may reach exercise levels more frequently during a day than in adults. When near water, for example, children tend to spend more time in the water than adults. This can lead to higher exposures of air pollutants, such as ozone, or to toxicants in water due to the inhalation of volatilized chemicals, ingestion of pool water, and dermal exposures.

Other behaviors characteristic of children such as overexposure to sunlight for sun tanning among adolescents, or drug and alcohol ingestion, can also lead to disproportionately higher risks. However, in this chapter, the central focus is on those risks that are completely involuntary, and that would be difficult to change without altering the normal behavior of a child.

The Unique Microenvironment of Children The children's microenvironment is different from that of adults from the very beginning of conception. The womb, where the fetus is exposed to environmental chemicals and other agents from previous exposures of the mother is a unique environment to this developmental stage. In fact, humans are among the living mammals that have the longest gestational period, resulting in a long in-utero exposure. Many pollutants are known to cross the placenta and enter the fetal circulation, resulting in fetal exposure to toxicants present in the mother. Once the baby is born, breastfeeding can also result in exposures to chemicals accumulated in breast milk.

Children spend a large proportion of their time indoors, and because of their smaller stature and because they crawl and play on the floor, they are more highly exposed to ground-level contaminants. In indoor environments, for example, formaldehyde exposure from carpeting would be higher at ground level than about 4 or 6 ft above, the breathing zone of adults. House dust or particles to which pollutants adhere accumulate near the floor both in the indoor and outdoor air. Further, outdoors where there is traffic, their breathing zone is close to the height at which car exhaust systems expel their fumes.

Indoor air pollutants can originate indoors, such as formaldehyde from carpeting, fumes from home cooking, and solvents from freshly painted rooms, or can enter the inside of homes through windows or open airways. Sometimes, pollutants may be carried inside by persons entering the indoor environment. For example, children whose parents work on farms where pesticides are applied can be exposed to pesticides carried inside in the clothes and skin of adults who walk into the house. Studies have found higher levels of pesticides in farmworker households in agricultural areas compared with those in non-farmworker households in the same area [9]. In addition, some studies have demonstrated that degradation of pesticides is slower in the indoor environment, so pesticide exposure persists longer [9].

Toxicokinetics Once exposed, the amount of toxicant being absorbed, its metabolism, and excretion from children's bodies may be very different from that of adults. Rates of absorption through the oral, dermal, and respiratory pathways vary between children and adults. For example, children have a larger lung surface area per kilogram of body weight than adults and, under normal breathing, breathe 50% more air per kilogram of body weight than adults [6]. Therefore infants potentially receive a greater internal exposure to airborne compounds on a body weight basis.

Children's absorption of chemicals through the gastrointestinal route is also different from that of adults. For example, the absorption of lead from the intestine was found to be 40 times higher for children than for adults [3]. Gastric pH is higher in newborns than in adults, thus causing differences in ionization and absorption of certain chemicals. The alkaline gastric pH in newborns and infants may enhance the absorption of certain compounds.

The skin surface area of children relative to body weight is greater for children than for adults, resulting in a higher potential dose absorbed through the skin of about three times greater for infants than for adults [10]. In addition, hydration of the skin in neonates is greater than in older children, which potentially could result in them having a greater absorption of some hydrophilic chemicals. The permeability of the epidermal is incomplete in the preterm infant, resulting in greater percutaneous absorption of chemical agents [5].

Once absorbed, the different body composition of children will result in a different distribution and accumulation of chemicals compared to adults. For example, the relatively larger extracellular fluid volume of the infant means somewhat greater dilution of water-soluble chemicals. However, the lipid-soluble substances would be distributed in a smaller volume of fat in infants relative to adults. The body composition in terms of water, fat, and protein changes from birth to adulthood resulting in corresponding changes in the bioaccumulation, metabolism, and excretion of contaminants from the body throughout development. Water content, for example, decreases rapidly during the first 6 months of life and then remains fairly constant. Body fat, on the contrary, increases rapidly up to 6 months and then decreases, accounting for similar percentages of body weight at ages 4 and 12 months [5]. This can result in changes in the concentration of chemicals in different tissues and organs that can override the child's ability to metabolize and excrete the chemical.

Metabolism and elimination rates are generally lower in neonates than in adults. Many metabolic pathways are not fully developed in the infant. Renal clearance is lower in neonates than in older children and adults for all classes of chemicals. These factors will lead to differences in the bioaccumulation and persistence of environmental agents in children's and adult's bodies and the concentration of the agents in specific organs. The maturation of these metabolic and elimination rates can result in the mobilization of chemicals, and higher exposures of the target organs or systems to the products of metabolism. In terms of excretion, certain life stages such as pregnancy, lactation, and also menopause, will result in the mobilization of chemicals from fat or bone stores, and the exposure of other organs or the developing fetus and lactating child to the toxicants that were accumulated in the mother.

Developmental Aspects of Children's Susceptibility: Critical Periods of Development

Effects During Gestation From conception to birth, the human organism advances rapidly through a complex set of developmental processes that culminate in the newborn. These development processes include cell division, organ formation, and growth as well and

functional development. During development, some biological processes occur only during certain stages of development and not in others, or they occur at a different rate in different developmental stages. For example, cell division in most organs takes place rapidly during early development and much more slowly at later stages. Other processes such as apoptosis, or programmed cell death, occur more widely during development and are less prominent during adulthood.

These biological processes need to be effectively coordinated and require the cellular and intercellular signaling systems to work correctly. Because of the complexity and speed at which these processes take place and the intricate relation between them, interference at the sequence of any of these processes can have damaging and irreversible effects. Exposure to environmental toxicants can have a completely different effect depending on whether it occurs at one developmental stage or another. In addition, damage due to environmental exposures may occur and manifest itself immediately, or may not appear until subsequent stages of development, after development is complete, including late adulthood.

Identifying and understanding these "critical" periods of unique susceptibility is essential to developing strategies to protect children from adverse health effects associated with environmental exposures. The following sections describe the current state of knowledge about the windows of susceptibility during childhood development.

Effects on Germ Cells Germ cells (sperm and egg cells) carry the genetic information from each parent that will provide the unique genetic blueprint for each child. In the male fetus, primordial germ cells develop in utero. From puberty to adulthood, these cells undergo cell division, mitosis, and meiosis, to produce mature sperm and continue to be produced from stem cells through adulthood. In females, primordial germ cells undergo mitosis and the first phase of meiosis during fetal life, and in women, mature oocytes are produced every month from follicular cells.

Environmental toxicants that harm germ cells can affect an adult's own fertility as well as the health of the offspring. In animal models, preconceptional carcinogenesis has been demonstrated for a variety of types of radiation and chemicals, with demonstrated sensitivity

for all stages from fetal gonocytes to postmeiotic germ cells [11]. Although this link in humans is not demonstrated, it is theoretically possible that environmental toxicants that harm germ cells can affect an adult's own fertility as well as the health of the offspring.

Results of environmental damage to germ cells may include reduced fertility later in life or offspring with congenital health problems [12, 13]. For example, men exposed to diethylstilbestrol (DES) in utero had lowered sperm count and increased frequency of abnormal sperm [14]. Women exposed to cigarette smoke during their mother's pregnancy had reduced fertility [13]. There is a substantial body of evidence demonstrating that exposures to environmental agents and medical radiation can injure germ cells in such a way as to cause increased incidence of cancer, particularly leukemia, among offspring of the exposed individuals. For example, paternal exposures to benzene have been linked to leukemia and lymphoma in children [15]. Animal studies support these findings [11].

Embryonic and Fetal Development During Pregnancy Several stages of embryonic and fetal development are susceptible to environmental harm. During development, gene expression is very active because a large number of genes are being "switched on" or "switched off" to control cellular activities. This high level of metabolic activity provides for a wide range of opportunities for environmental agents to interfere with cell development and growth.

Environmental toxicants may interact directly with DNA (e.g., alkylating agents) and disturb gene expression or may interact with the products of gene expression, such as enzymes and control molecules. A toxicant that interferes with gene expression may prevent the synthesis of enzymes necessary for toxicant metabolism, resulting in the accumulation of the toxicant in the body. DNA activation by a chemical may result in excessive synthesis of enzymes that catalyze the bioactivation of a toxicant (production of a toxicant metabolite more toxic than the compound originally present).

Chemicals can interfere with the activation or inactivation of genes that occur during early fetal life and that may be essential for the protection of the organism to external or internal harmful processes. For example, interference with genes involved in DNA repair, such as

p53, a tumor suppressor gene important for DNA repair, may result in enhanced vulnerability to specific toxicants during development. Studies with transgenic mice that are missing this gene have shown an increased sensitivity of mice fetuses to benzo[a]pyrene exposure and an increased death rate when exposed to the chemical during gestation [16]. Damage to p53 in humans could likewise increase sensitivity to agents that damage genetic material.

If a toxicant interferes with cell differentiation, cells may not reach their specific form and function necessary for their final role in the body, and organ function may be compromised. Also, undifferentiated cells may be more vulnerable than differentiated cells to toxic effects. Some chemicals such as ethanol [17] and TCDD [18] have been demonstrated to affect specific types of undifferentiated cells.

Apoptosis or programmed cell death is a critical biological process for healthy development. Apoptosis involves the removal of certain cell types when they are no longer necessary. In some instances, one type of cell is succeeded by another during a specific developmental period. Apoptosis is involved, for example, in the elimination of cells in the immune system that, if they survived, could cause autoimmune disease [19]. Apoptosis is also critical in the development of the nervous system, where phases of cell proliferation alternate with phases of apoptosis on the basis of the progression of neuronal development [20] and remains active through the postnatal period because of ongoing nervous system development.

Normal patterns of apoptosis may be altered through altered gene expression or failure of signaling mechanisms resulting from environmental exposures. Certain autoimmune lympho-proliferative diseases and certain cancers have been related to the disruption of normal patterns of apoptosis. For example, Wilms' tumor, a relatively common childhood cancer, may arise from the transformation by postnatal exposures of renal stem cells that fail to disappear 4–6 weeks prior to birth [21].

Neuronal migration is an important process in nervous system development and its alteration may result in irreversible damage. For example, schizophrenia is thought to result, in part, from abnormal neuronal migration [22], but the role of prenatal exposures to environmental agents in causing

this disease is not clear. Exposures to ionizing radiation and methylmercury, for example, have been shown to affect the migration of neurons during development [20, 23].

During the period of organ development, which occurs (varying according to organ system) between the 3rd and 16th week, disruption of development can disrupt the large-scale structure of organs, often resulting in physical malformations (congenital anomalies). The best known example of such gestational damage is exposure to diethylstilbestrol (DES). DES caused genital anomalies among male children born of women who took the medication before the 11th week of gestation twice as often as among those who were exposed later in gestation [14].

Other effects such as low birth weight, pregnancy complications, or late fetal death have been shown to be a result of environmental exposures during later stages of prenatal development [24]. Disinfection by-products have been linked to the risk of spontaneous abortion for some time. There is now fairly consistent evidence for associations between early and late fetal deaths and indices of transplacental exposure to disinfection by-products [25–27]. Maternal smoking during pregnancy increases the risk of pregnancy loss, stillbirth, and infant mortality [28].

Development During Childhood Several organs and systems continue to grow and develop during childhood and in some cases almost until adulthood. For example, neuron migration, cell proliferation, and synapse formation are very active until 3 years of age, and myelination continues until adolescence [29] and possibly well into adulthood [20].

The immune response is also immature at birth and develops during infancy and childhood until about 1 year of age, while establishment of immunologic memory is not fully established until 18 years of age [30]. Exposure to environmental agents during early childhood may affect immune system development and may contribute to the development of certain diseases such as asthma and cancer later in life.

Physical growth and maturation of organ systems continues through adolescence. The process of sexual maturation is accompanied by complex interactions between the central nervous system and hormone-secreting organs, which can be affected by

environmental exposures. For example, the risk of breast cancer has been found to be greater among women who were exposed to radiation before 20 years of age [31].

Cellular Metabolism and Biotransformation Many important metabolic and biotransformation processes are poorly developed in the fetus, and full metabolic activity is not fully developed until after childbirth. Metabolism can increase or decrease the toxicity of a chemical, depending on the metabolic products of the chemical and pathway involved. Metabolism may also make elimination from the body easier or harder, although the most common metabolic pathways usually render chemicals more hydrophilic and thus, more easily excreted. In some cases, the adult biotransformation of a certain chemical may consist of a bioactivation pathway that makes the compound more hazardous than the one originally present. The absence of a metabolic pathway may result in the bioaccumulation of the chemical in the body and a later bioavailability and disposition to exert its toxic effects. Immaturity could be an advantage if the activation pathway is not present in the fetus or child and there is an alternate pathway for the toxicant to be metabolized. However, according to [32], given the primary evolutionary function of detoxifying and eliminating potentially toxic chemicals, immature or underdeveloped metabolic pathways are likely to render infants and children more sensitive to common environmental contaminants.

Major Groups of Pollutants to Which Children are Exposed

Heavy Metals

Heavy metals are natural elements that have been extracted from the earth and used in human industry and products for centuries. As a consequence of human activity, concentrations of heavy metals in air, water, and surface soil today are hundreds of times higher than in the preindustrial era. Some metals are naturally found in the body and are essential to the functioning of critical enzyme systems. Iron, for example, prevents anemia, and zinc is a cofactor in over 100 enzyme reactions. Magnesium and copper are other familiar metals that, in minute amounts, are necessary for

proper metabolism to occur. The body has need for approximately 70 trace elements, but there are others, such as lead, mercury, aluminum, arsenic, cadmium, and nickel, that have no roles in human physiology and can be toxic at even trace levels of exposure. Nutritionally, heavy metals can compete with nutrient elements, such as the case of lead, which is stored in the bones in the place of calcium.

Metals are notable for their wide environmental dispersion, their tendency to accumulate in select tissues, and their overall potential to be toxic at even relatively minor levels of exposure. In general, heavy metals are systemic toxins with specific *neurotoxic*, *nephrotoxic*, *fetotoxic*, and *teratogenic* effects. Heavy metals can directly influence behavior by impairing mental and neurological function, influencing neurotransmitter production and utilization, and altering numerous metabolic body processes.

Exposure to heavy metals can occur through drinking water, air, or ingestion of heavy metal-contaminated soil. The amount that is actually absorbed from the digestive tract can vary widely, depending on the chemical form of the metal and the age and nutritional status of the individual. Once a metal is absorbed, it distributes in tissues and organs. Excretion of metals typically occurs through the kidneys and digestive tract, but they tend to persist in some storage sites, like the liver, bones, and kidneys, for years or decades.

Lead is one of the best known heavy metals in terms of its toxicity. Exposure to lead can occur in the prenatal stage through the placenta, and in infants through the mother's milk and the water used in milk formula [3]. During pregnancy, body stores of lead may be mobilized and transferred from the mother to the fetus [33]. Behavioral characteristics of children later on, such as the hand-to-mouth behavioral pattern of 1–3 year olds, can result in high exposure and internal levels of lead. Lead paint is a major source of environmental exposure for children who ingest flaking paint, paint chips, and weathered powdered paint (mostly from deteriorated housing units in urban areas). Lead can leach into drinking water from lead-based solder used in water pipes. Lead also leaches into foods or liquids stored in ceramic containers made with lead glazing, which is still used in some countries.

The absorption of lead from ingestion of lead-contaminated water is higher for children than for adults, so that for a given level of exposure, the resultant internal dose is higher in children than in adults [3]. Children are also more sensitive than adults to the toxicological effects of lead at a given internal exposure level. The lowest observed adverse effect levels (LOAELs) for several health endpoints occur at lower blood lead levels in children than in adults. The most sensitive targets for lead toxicity are the developing nervous system, the hematological and cardiovascular systems, and the kidney. There appear to be no safe exposure “thresholds” for lead or for other metals in early development.

Mercury is a ubiquitous heavy metal of both natural and anthropogenic sources. Mercury occurs in both inorganic and organic forms, and it is most hazardous in its organic form of methylmercury. The nervous system is very sensitive to all forms of mercury. Methylmercury and metallic mercury vapors are more harmful than other forms, because methylmercury can cross the blood brain barrier. Methylmercury in the marine and freshwater environment is absorbed by fish and shellfish and bioaccumulates in the food chain. Increased risk is of particular concern in children and in populations that have an increased dietary exposure to fish.

Arsenic occurs naturally in the environment and in some areas of the world is a natural contaminant of underground water that is used as drinking water. It is also an anthropogenic contaminant. Once absorbed into the body, arsenic undergoes some accumulation in soft tissue organs such as the liver, spleen, kidneys, and lungs, but the major long-term storage site for arsenic is keratin-rich tissues, such as skin, hair, and nails.

Cadmium is another chemical that is toxic to adults, although it has not been extensively studied in children. The US Department of Health and Human Services has determined that cadmium and cadmium compounds are known human carcinogens. Cadmium has been linked to diminished kidney function lung disease, chronic bronchitis, and lung, kidney, and prostate cancers. In the USA, smoking is the primary source of cadmium exposure, although high levels of cadmium can also be found in organ meats, shellfish, and vegetables.

Pesticides

Pesticides are substances that are used to prevent, repel, or destroy pests – organisms that compete for food supply, adversely affect comfort, or endanger human health (FIFRA 1996). More than 20,000 pesticide products with nearly 900 active ingredients are registered for use as insecticides, miticides, fumigants, wood preservatives, and plant growth regulators. It cannot be denied that pesticides have beneficial economic and also public health impacts. Pesticide usage helps improve human nutrition through greater availability, longer storage life, and lower costs of food. It also reduces human labor requirements and attendant risks of injury. Pesticides also assist in the control of food-borne and vector-borne diseases, such as malaria, which kill millions of persons in the world. Pesticides also pose human health concerns because they are toxic substances and widely spread in the environment. Although the toxic mechanisms on targeted pest species are well characterized, the potential for adverse health effects in humans is not fully known.

Pesticides are composed of several classes of chemicals with different mechanisms of action. Most insecticides work by interfering with nervous system function. Organophosphates, which account for approximately one-half of the insecticides used in the USA, and carbamates, which are widely used in homes and gardens, inhibit the activity of acetyl cholinesterase at nerve endings, resulting in an excess of acetylcholine in the synapsis and a depolarizing blockage of neural transmission. The effects of carbamates are readily reversible and of shorter duration. Organochlorines, such as dichlorodiphenyltrichloroethane (DDT) and lindane, interfere with nerve cell membrane cation transport, resulting in neural irritability and excitation of the central nervous system. Herbicides, including the chlorophenoxy compounds 2,4 D and 2,4,5 T are primarily irritative to the skin and respiratory tract during acute exposures and work by different mechanisms. Some substances, such as paraquat, are highly corrosive and can cause multisystem injury and progressive pulmonary failure [34].

Arsenical pesticides, such as copper chromium arsenate, have been used, until recently, as wood preservatives to prolong the useful life of exterior wooden structures. These compounds cause central nervous system depression at sufficient doses.

Pesticides are ubiquitous in the environment. They are found in food, water, homes, schools, workplaces, lawns, and gardens. They are present in soils that have been spread with pesticides or where pesticides from adjacent agricultural areas have drifted, and may reach water supplies from agricultural runoff. In the USA alone, more than 0.45 billion kilograms of pesticides are applied each year, in agriculture, in homes and gardens, and in schools and hospitals. In developing countries, many highly toxic and biologically persistent pesticides such as parathion, DDT, and paraquat, which are no longer permitted in many developed countries, are still in wide use and result in chronic exposures and acute, too often fatal poisonings of thousands of young children each year.

Most children in the world are exposed to some degree to pesticides. Children in rural and agricultural areas and especially children whose parents are farm-workers or pesticide applicators are at highest risk of having increased exposures to pesticides. Pesticides may reach their homes due to the drifting of pesticides that are applied to the ground through aerial spraying. Children may work or play near their parents in the fields where pesticides have been used. Parents who work with pesticides may bring pesticides to their homes, impregnated in their clothes and bodies. In countries where residential housing with gardens and lawn predominate, homes and garden pesticide use may result in significant levels of exposure [34].

Exposure of children to pesticides may occur through inhalation, ingestion, and dermal absorption. Ingestion of pesticides occurs either through accidental exposure due to pesticides stored in food containers (i.e., soft drink bottles), or through ingestion of pesticide-treated foods, particularly fruits and vegetables. Foods grown in pesticide-contaminated soils and fish from pesticide-contaminated water can also carry significant amounts of pesticides. Children may also ingest pesticides adhered to the surface of toys or other objects through their hand-to-mouth behavior. The potential of dermal exposure of children to pesticides is higher than that of adults because of their relatively large body surface area and extensive contact with lawns, gardens, and floors by crawling and playing on the ground.

Prenatal and early childhood exposures are of special concern because of the susceptibility of the

developing organ systems to pesticides (the central nervous system in particular) as well as behavioral, physiological, and dietary characteristics of children. Breast-feeding infants may ingest pesticides or pesticide metabolites present in the breast milk. The quantity of pesticide that is passed to the infant via breast milk is influenced by many variables such as maternal age and parity, maternal body burden of the chemical, and breast-feeding patterns. As infants are weaned and progress to solid foods, they consume, per unit of body weight, proportionally more fruit and more fruit juice than adults. The NAS in 1993 reported that children's dietary exposures to pesticides differed from adults both quantitatively and qualitatively and questioned the protection provided to infants and children from pesticide tolerances in effect at the time. The report estimated that 50% of lifetime pesticide exposure occurs during the first 5 years of life [2].

Environmental Tobacco Smoke

Environmental tobacco smoke (ETS), also known as second-hand smoke, is exhaled smoke and sidestream smoke emitted from the burning of the tip of the cigarette. The inhalation of ETS is known as "involuntary smoking" or "passive smoking." ETS contains more than 4,000 different chemical compounds, many of which are toxic. In 1992, the US Environmental Protection Agency (EPA) declared ETS as a Group A carcinogen.

The effects of passive smoking begin in utero, where constituents of tobacco smoke, such as PAHs, nicotine, and carbon monoxide, cross the placenta and are concentrated in the fetal circulation [35]. Children are also exposed during childhood if any of the parents smoke.

Persistent Organic Compounds: PCBs, Dioxins, and Related Organohalogenes

Polychlorinated biphenyls are synthetic compounds with two linked phenyl rings and variable degrees of chlorination. They have been used for many years because of their thermal and chemical stability. They are nonvolatile, hydrophobic oils that are not easily biotransformed in the environment or metabolized by the human organism, so they are very persistent in the environment, and bioaccumulate in the food chain

and in the fat compartment of the human body. Although they have been banned in the USA for more than 30 years, they are still widely present in the environment. They have been found in wildlife, human tissue, and human milk. Polychlorinated dibenzofurans (PCDFs) are partially oxidized PCBs that appear as contaminants of PCBs. Polychlorinated dibenzodioxins (PCDDs), commonly referred as dioxins, are formed during paper bleaching and waste incineration. One dioxin congener, 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD), is considered to be the most toxic synthetic chemical known.

The most significant source of exposure is contaminated food, particularly fish at the top of their trophic chain. Of greatest concern are the populations that consume high amounts of fish. Because PCBs and PCDFs are not metabolized or excreted, they can accumulate in the fat tissue of the body, as well as in human milk, resulting in high exposures of the developing fetus and the newborn.

Disinfection By-products (DBPs)

Disinfection by-products (DBPs) form when disinfectants are added to drinking water and react with naturally occurring organic matter. Chlorine, the most widely used primary disinfectant, reacts with naturally occurring organic matter to form a range of unwanted by-products such as the trihalomethanes (THMs) which include chloroform, bromodichloromethane (BDCM), chlorodibromomethane (DBCM), and the haloacetic acids (HAAs), such as monochloroacetate, dichloroacetate, and trichloroacetate.

Exposure to DBPs occurs through ingestion of water or through inhalation and absorption during showering, bathing, and swimming. While there is some concern that these chemicals may pose a health risk, the potential risks arising from not treating drinking water are considerably greater, and the disinfection of water should never be compromised as a result.

Environmental Threats to Children on Specific Organ Systems

Nervous System

The developing nervous system is more susceptible than the adult brain to the disrupting effect of toxic

chemicals [5]. The lengthy period of brain development and the extensive number of processes needed to take place contribute to the susceptibility of the developing nervous system. In the 9 months of pregnancy, the human brain and spinal cord must develop from a thin strip of cells along the back of the embryo into a complex organ comprised of billions of precisely located, highly interconnected, and specialized cells. Brain development requires that neurons move along precise pathways from their points of origin to their assigned location, that they establish connections with other cells near and distant, and that they learn to intercommunicate. Each connection between and among neurons must be precisely established at a particular point in development, and redundant connections need to be pruned away through programmed cell death, apoptosis. All these processes must take place within a tightly controlled time frame, in which each developmental state must be reached on schedule and in the correct sequence.

Critical windows of vulnerability, which exist only in the 9 months of pregnancy and to a lesser extent in early childhood, are unique to early brain development. They have no counterpart in adult life. Exposure to toxic chemicals during these windows of vulnerability can cause devastating damage to the brain and nervous system.

Any toxic or other environmental exposure that interferes with the tightly orchestrated sequence of events involved in brain formation is likely to have profound effects on intellect, behavior, and other functions. If a developmental process in the brain is halted or inhibited, if cells fail to migrate the proper sequence to their assigned locations, if synapses fail to form, or if pathways are not established, there is only limited potential for late repair, and the consequences can be permanent [36].

Environmental toxicants can affect both the structural and functional development of the nervous system. Depending on the developmental stage at which exposure occurs, sensory development, intelligence, or behavior will be affected differentially. While early-developing neural systems have been considered the most vulnerable to chemical insult, scientists have called attention to the importance of chemical exposure that occurs late in childhood, as it has been recently suggested that behavioral and physiological

foundations of cognition continue to develop during childhood and adolescence [5, 37].

Exposure to environmental toxicants such as lead, methylmercury, and certain pesticides and PCBs even at very low levels have been shown to produce neurobehavioral (functional) deficits, and increased susceptibility to neurodegenerative diseases much later in life [38]. Of critical concern is the possibility that developmental exposure to neurotoxicants may result in an acceleration of age-related decline in function that could lead to Parkinson Disease, Alzheimer Disease, and other forms of brain degeneration. This concern is compounded by the fact that developmental neurotoxicity that results in small effects on the individual at a particular stage can have a profound societal impact when considered in the whole population or across the life span of the individual [36]. For example, a five-point decline in average population intelligence (IQ score) can result in reduction by more than 50% in the number of children with superior IQ (>130) and concomitant doubling in the number of children with IQ scores in the retarded range (<70). Behavior is also disrupted, sometimes permanently, by such exposures.

The following sections will give a short overview of some neurotoxicological effects of the best known chemicals for their effect on the developing nervous system. Some toxicants such as ethanol have been excluded, as the focus of this chapter is on exposures to toxicants to which the majority of children are exposed.

Lead Lead is one of the best studied toxicants and one of the few pollutants for which susceptibility of children has been clearly established. The best known health effects on children are its neuropsychological effects although other effects have also been studied and are partially documented.

The neurotoxic effects of lead in children have been extensively studied. One distinctive characteristic of the research findings on lead neurotoxicity over the past 2–3 decades is that it has led to a progressive decline in the LOAELs (lowest observed adverse effect levels). The early finding in the 1970s by Landrigan et al. [39] and Needleman [40] that low-dose exposure to lead is associated with a significant decrease in intelligence quotient (IQ) has been confirmed in many

studies. The recommended action level of blood lead levels established by the Centers for Disease Control is 10 $\mu\text{g}/\text{dL}$ of blood. However, epidemiological studies during the last decade have found strong evidence for cognitive deficits among school-aged children at blood lead levels below the current CDC action level [41, 42]. Further, in a pooled analysis of seven prospective longitudinal studies, the average IQ deficit associated with an increase in concurrent blood lead concentration from <1 to 10 $\mu\text{g}/\text{dL}$ was about threefold that for an increase from 10 to 20 $\mu\text{g}/\text{dL}$ [42]. Birth cohort studies have shown inverse dose-response relationships between transplacental lead exposure indices and central auditory processing indices among infants and children at maternal blood lead levels below 10 $\mu\text{g}/\text{dL}$ [43].

Research has also demonstrated a link between developmental lead exposure and behavioral outcome. In a prospective study, the behavior of lead-exposed children at 8 years of age was significantly related to tooth dentine levels [44], suggesting that social and emotional difficulties correlate with lead exposure. In another prospective, longitudinal study, both prenatal and postnatal lead exposure was related to antisocial and delinquent behavior in adolescents [45]. It is generally accepted in the scientific and medical community that the adverse neurobehavioral consequences of lead are not reversible and remain in place across the life span [41]. Further, the possibility that a threshold level for the effects of lead does not exist has been suggested [42].

Mercury Methylmercury is a well-established neurotoxicant that can cause serious adverse effects on the development and functioning of the human central nervous system, especially when exposure occurs prenatally. The well-known episodes of community-wide poisoning in Japan and Iraq revealed the particular sensitivity of the fetus to the toxic effects from mercury exposure. In these communities, pregnant women exposed to methylmercury and who themselves had no or minimal symptoms, had babies with devastating neurological handicaps, including delayed attainment of developmental milestones, blindness, deafness, and cerebral palsy.

At levels of exposure lower than those encountered in the Minamata Bay, there is limited epidemiological

evidence of neuropsychological effects. A birth cohort of 1,000 children was established at the Faroe Islands in 1986–1987, and the methylmercury exposure was determined from the mercury concentration in the cord blood [46]. More than 90% of these children were then examined at age 7 years. Neuropsychological effects in the areas of language, attention, and memory and to a lesser extent in visuospatial and motor functions were observed [46]. In Brazil, cross-sectional studies of Amazonian children aged 7–12 years show mercury-associated effects in agreement with the Faroe findings [47]. However, a large cohort study conducted in the Seychelles did not reveal consistent associations between perinatal methylmercury exposure indices and developmental milestones or neuropsychologic test scores up to age 5 years old [48].

Some studies have found an association between central auditory processing deficits and current childhood hair mercury levels as low as 5 $\mu\text{g}/\text{g}$ [49]. These findings are supported by animal studies, showing that transplacental or postnatal methylmercury exposure affects auditory systems at the cortical level [50].

The important question is to which degree these findings relate to fish-consuming populations in general. Although this question cannot be answered with any confidence at this time, it looks like the recommended one to two fish meals per week during pregnancy would be very unlikely to cause any risk to the fetus, unless the seafood is severely contaminated. The National Academy of Sciences recommended that a limit of about 0.1 $\mu\text{g}/\text{kg}$ of body weight per day should not be exceeded by pregnant women [51].

PCBs, Dioxins, and Related Organohalogenes

Polychlorinated biphenyls are known to interfere with thyroid hormones, some of which are critical for normal brain development, and this effect of PCBs on the thyroid function of the newborn has been postulated as the mechanism of action of neuropsychological effects [52].

The earliest evidence of PCB-related neurotoxicity comes from the poisoning episodes in Japan (Yusho) and Taiwan (Yucheng) where people became ill after ingesting rice oil that was highly contaminated with PCBs. Infants born to mothers who consumed PCB-contaminated rice oil during pregnancy were at increased risk for low birth weight, abnormal brown

pigmentation of the skin, and clinical abnormalities of the gingival, skin, nails, teeth, and lungs [53]. In addition, children of both cohorts had various neurobehavioral deficits such as delayed attainment of developmental milestones, lower scores on intelligence tests, and higher activity levels [54]. Children were followed and examined 6 years later and showed persistent behavioral abnormalities and ectodermal defects [55].

Many studies after the Taiwan event provide evidence of neuropsychological effects of PCBs [56]. The evidence indicates that PCBs cause neurobehavioral deficits in children who are exposed prenatally, while the evidence of effects of exposure to background levels commonly found in the general population is not conclusive. Studies support the association between low-level transplacental PCB exposure and childhood deficits in cognitive, psychomotor, memory, language, and attention functions [56, 57]. None of the studies up to date has been able to document an adverse effect of PCBs exposure from breast-feeding. It has been postulated that the larger fat compartment of the nursing infant than the fetus may make it possible to dilute somewhat the lipid-soluble contaminants absorbed through human milk. Thus, despite the occurrence of these contaminants in human milk, the advantages of breast-feeding for 4–6 months apparently override any limited neurotoxin damage due to the contaminants. Still, of particular concern would be the populations who frequently eat contaminated fish or who reside in contaminated areas and may still be exposed at levels that have been associated with adverse effect.

Pesticides Current knowledge about the neurotoxicological effects of pesticides is limited, despite the extensive knowledge available of the mechanisms of pesticide toxicity in animals. One observational study in children from a region in Mexico with intensive pesticide use found a variety of developmental delays compared with otherwise similar children living in a region where the population had not adopted a pesticide-based agriculture. The children were similar in growth and physical development, but significant delays were noted among the exposed children in physical stamina, gross and fine hand-eye coordination, short-term memory, and ability to draw a human

figure [58]. However, these conclusions have been questioned, as pesticide levels were not reported for the individual children who received neuropsychological testing.

Fetal and neonatal animals are often more sensitive than adults to the neurotoxic effects of some organophosphates [2], and levels presumed to be nontoxic in adults may not be adequately protective of the developing organisms. For example, the young rat is deficient in the actions of two enzymes that detoxify chlorpyrifos, and young rats have increased sensitivity to chlorpyrifos toxicity [59]. In rats, chlorpyrifos appear to affect cholinergic function [60]. Infants may be particularly vulnerable to reductions in brain cholinesterase given that acetyl choline plays an important role in normal brain development and plasma and erythrocyte (and therefore probably brain) cholinesterase do not reach adult values until 6–12 months of age [61].

Animal studies suggest that exposure to pesticides such as DDT and its metabolites DDE and dichlorodiphenyldichloroethane at the levels found in the environment affect the developing brain. Ten-day old mice treated once with DDT showed behavioral changes compared with controls when tested at 4 months of age. These effects occur only after DDT dosing at 10 days of age and not after similar dosing at 3 or 19 days of age [62].

In spite of the absence of direct conclusive evidence of the neuropsychological effects of pesticides, the known subclinical effects observed in adults through neuropsychological testing, and what is known on the mechanism of action of pesticides and susceptibility of the developing brain, has led many investigators to infer that chronic low-dose exposure to certain pesticides might pose a potential hazard to the health and development of infants and children. Other investigations have concluded such inferences can be neither supported nor refuted at the present time.

Endocrine System

The endocrine system is one of the body's main communication networks and is responsible for controlling and coordinating numerous body functions. Hormones are first produced by the endocrine tissues,

such as the ovaries, testes, pituitary, thyroid, and pancreas, and then secreted into the blood to act as the body's chemical messengers. They direct communication and coordination among other tissues throughout the body, and they exert a powerful influence over growth and development of every organ system in the body.

The endocrine system regulates metabolic, nutritional, reproductive, and behavioral processes, as well as growth, responses to stress, and the function of the digestive, cardiovascular, renal, and immune system. Programming of endocrine set points is a unique aspect of endocrine system development that takes place during fetal/neonatal development, and exposure to toxicants during this critical period of programming can result in permanent abnormalities in endocrine function [5]. Disruption of endocrine function can have severe health consequences in adults, and exposures that interfere with the development of the endocrine system during early life stages can have even more far-ranging consequences [63].

Endocrine disruptors are natural compounds or man-made chemicals that may alter the production or activity of hormones of the endocrine system leading to adverse health effects. Although there is limited scientific information on the potential adverse human health effects, concern arises because endocrine disrupting chemicals, while present in the environment at very low levels, have been shown to have adverse effects in wildlife species, as well as in laboratory animals at these low levels. Many of these chemicals have been linked with developmental, reproductive, neural, immune, and other problems in wildlife and laboratory animals. The potential adverse effects of EDCs include neurodevelopmental and neurobehavioral abnormalities, reproductive disorders such as declined fertility, immune impairment, and certain hormone-related cancers [1].

An example of a phthalate is di(2-ethylhexyl) phthalate (DEHP). DEHP is a high-production-volume chemical used in the manufacture of a wide variety of consumer food packaging, some children's products, and some polyvinyl chloride medical devices. Recently, an independent panel of experts assembled by the National Toxicology Program (NTP) found that DEHP may pose a risk to human development, especially critically ill male infants.

In 2000, an independent panel of experts convened by the NIEHS and the National Toxicology Program (NTP) found that there was "credible evidence" that hormone-like chemicals can affect test animals' bodily functions at very low levels – well below the "no effect" levels determined by traditional testing. Although there is little evidence to prove that low-dose exposures are causing adverse human health effects, there is a large body of evidence in experimental animals and wildlife suggesting that endocrine disruptors may cause [1]:

- Reductions in male fertility and declines in the numbers of males born
- Abnormalities in male reproductive organs
- Female reproductive diseases including fertility problems, early puberty, and early reproductive senescence
- Increases in mammary, ovarian, and prostate cancers

Endocrine disruptors may interfere with the endocrine system through several mechanisms. Some mimic or partially mimic occurring hormones in the body like estrogens and androgens and thyroid hormones, potentially producing overstimulation. Another group of natural and synthetic substances interferes with the hormones at receptors. Substances that compete with a hormone at the receptor, and imitate its effect are called agonists, those that block the receptor are antagonists. Other chemicals interfere, or block the way natural hormones or their receptors are made or controlled, for example, by blocking their metabolism in the liver. Environmental chemicals with estrogenic activity are probably the most well studied; however chemicals with antiestrogen, androgen, antiandrogen, progesterone, or thyroid-like activity have also been identified.

A wide range of substances are thought to cause endocrine disruption. Chemicals that are known endocrine disruptors include diethylstilbestrol (DES), dioxin and dioxin-like compounds, PCBs, DDT, and some other pesticides. Some chemicals, particularly pesticides and plasticizers, such as Bisphenol A are suspected endocrine disruptors based on animal studies.

Phthalates, a class of chemicals that soften and increase the flexibility of polyvinyl chloride plastics, are considered to be endocrine disruptors on the basis of extensive animal research plus emerging studies in humans. An example of a phthalate is

di(2-ethylhexyl) phthalate (DEHP). DEHP is a high-production-volume chemical used in the manufacture of a wide variety of consumer food packaging, some children's products, and some polyvinyl chloride medical devices. Recently, an independent panel of experts assembled by the National Toxicology Program (NTP) found that DEHP may pose a risk to human development, especially critically ill male infants.

Research shows that endocrine disruptors may pose the greatest risk during prenatal and early postnatal development when organ and neural systems are developing. In animals, adverse consequences, such as subfertility, premature reproductive senescence, and cancer, are linked to early exposure, but they may not be apparent until much later in life.

There is some evidence that endocrine disruptors may not only impact the individual directly exposed but also future generations. It has been found that animals exposed to low doses of the natural human estrogen estradiol, or the environmental estrogen bisphenol A (BPA) (a chemical used in great quantities in the production of polycarbonates and epoxy resins), during fetal development and estradiol as adults were more likely to develop a precursor of prostate cancer than those who were not exposed [64]. This suggests that exposure to environmental and natural estrogens during fetal development could affect the way prostate genes behave, and may lead to higher rates of prostate disease during aging. It has also been shown that the adverse effects of diethylstilbestrol in mice can be passed to subsequent generations even though they were not directly exposed [64]. The increased susceptibility for tumors was seen in both granddaughters and grandsons of mice who were developmentally exposed to DES [65]. One study found that endocrine disruptors caused fertility defects in male rats that were passed down to nearly every male in subsequent generations [66]. These intergenerational effects may be the consequence of epigenetic changes caused by endocrine disruptors.

There is concern that alterations in thyroid hormone signaling by endocrine disruptors during fetal and neonatal development could disrupt central nervous system development. There have been several epidemiological studies of the neurobehavioral effects of in utero exposure to PCBs that have been mentioned in the section on nervous system effects. Thus far,

few studies have directly linked neurobehavioral effects of exposure to polyhalogenated hydrocarbons to disruption of thyroid hormone signaling. One example is a study that showed that low-frequency hearing loss caused by developmental exposure in rats to PCBs could be partially reversed by replacement of T4 [67]. In animals, many polyhalogenated aromatic hydrocarbons such as dioxins and PCBs alter thyroxine levels via an increased metabolism and excretion of these hormones [52].

Perchlorates are another class of environmental chemicals that affect thyroid function via inhibition of iodine uptake by the thyroid gland, reducing T4 and T3 synthesis. There has been a concern that contamination of drinking water with perchlorates from industrial sites would suppress fetal thyroid hormone synthesis, disrupting central nervous system development. Increased rates of congenital hypothyroidism have been found in communities with detectable perchlorate levels in the drinking water [68], although others have not detected such elevated rates [69].

Evidence from epidemiological studies demonstrates that exposures during early life can result in greater susceptibility to diabetes and obesity later in life [70]. In the course of development of the pancreas and pancreatic function, for example, several windows of susceptibility have been identified [71]. Studies about the Dutch Famine Winter have shown that poor maternal nutrition, especially during the last trimester of pregnancy had an impact on glucose tolerance and insulin resistance, and that in terms of obesity, those individuals born to mothers who were exposed to the famine around the first half of pregnancy were more obese than those of mothers exposed during the last trimester [72]. A number of subsequent studies have documented the role of the intrauterine environment, and low birth weight in the development of diabetes [71, 73] and other metabolic disorders.

Reproductive System

The development of the reproductive system is a long process that takes place from the beginning of gestation, when organ development starts to take place, to maturation of the reproductive system during puberty. This long developmental period provides for several windows of susceptibility. Adverse effects of

reproductive toxicants can become manifest at birth (e.g., hypospadias and cryptorchidism in humans), in puberty (as delay or precocity), or in adulthood (e.g., infertility, alterations in accessory sex organs, disturbances in pregnancy maintenance, endometriosis, or premature reproductive senescence) [74].

The developmental susceptibility of the reproductive system was clearly demonstrated through the diethylstilbestrol (DES) exposure. In utero exposure of men to DES has been linked to increased incidence of meatal stenosis, epididymal cysts, testicular hypoplasia, cryptorchidism, micropallus, and sperm abnormalities. In females, DES exposure resulted in adenosis, clear cell adenocarcinoma, and structural defects of the cervix, vagina, uterus, and fallopian tubes [75].

Environmental tobacco smoke has been associated with decreased fecundity and earlier menopause in women smokers [76]. Women whose mothers smoked while pregnant also had reduced fecundity compared with women whose mothers did not smoke [13]. In men whose mothers smoked tobacco during pregnancy, reduced semen quality, smaller testis size, and reduced fecundability odds ratios have been observed [77].

Phthalates such as diethylhexyl phthalate (DEHP) are developmental toxicants in experimental animals. The organ system most sensitive to phthalates is the reproductive tract of immature males [78]. In utero exposure of male rats to some phthalate esters results in changes in the reproductive tract, such as decreased anogenital distance, hypospadias, cryptorchidism, disturbed development of prostate, epididymis, vas deferens, and seminal vesicles, retained nipples and decreased sperm production [79–82]. In humans, similar dysgenetic changes in the histology of the testis have been found in patients with testicular cancer, subfertility, or cryptorchidism [83, 84]. It has been hypothesized that all these human disorders (testicular germ cell cancer, cryptorchidism, hypospadias, and low sperm counts) have common origins in fetal life and that they all represent different symptoms testicular dysgenesis syndrome [87].

Neonatal Mortality, Growth Restriction, and Birth Defects

Birth defects are normally defined as structural problems in the newborn, attributable to faulty

development or deformation, but defects in function, metabolism, or body chemistry that lead to physical or mental problems or to death may also be considered birth defects. The broader term “developmental disorders” is generally used when considering all effects observed on the conceptus from fertilization to sexual maturity. Developmental disorders include both structural birth defects and functional defects, such as blindness, deafness, or neurobehavioral disabilities.

The majority of birth defects are considered the result of multiple environmental and/or genetic causes acting together. Environmental toxicants studied for their relation with birth defects include maternal smoking and alcohol use, pesticides, disinfection by-products, plastics and plastic components, solvents, metals and numerous air pollutants [88]. Other environmental causes such as maternal diseases (e.g., rubella), and use of pharmaceuticals (e.g., valproic acid, an anticonvulsant, and mood stabilizers), have served to document the potential environmental role in developmental birth defects.

Many epidemiological studies have attempted to evaluate whether some chemical exposures are linked to increased rates of spontaneous abortions. For example, occupational exposure of the mother to organic solvents has been associated with spontaneous abortion in several studies [89, 90]. A review of six occupational studies found suggestive evidence for an association between toluene exposure and spontaneous abortion, although most workers were simultaneously exposed to multiple chemicals, which may have confounded the interpretation of results. According to this review, spontaneous abortion has generally not been observed as a major problem among highly exposed women who abuse toluene during pregnancy.

Many studies have investigated the effects of *disinfection by-products* (DBPs) on the developing fetus and children. Review of epidemiologic studies reveal that there is fairly consistent evidence for associations between early and late fetal deaths and indices of transplacental DBP exposure [25, 26, 89]. Associations between late fetal deaths and drinking water chloroform, bromodichloromethane, and total THM concentration at levels below the Canadian drinking water guideline of 100 µg/L have been found in Canada [90]. In England, increased risks of late fetal deaths among women living in regions with THM concentrations

above 30 µg/dL [91]. These findings are supported by animal studies, where high prenatal maternal exposure to chloroform, bromodichloromethane (BDCM), haloacetonitriles, or haloacetic acids produced fetal resorption and reduced fetal survival [26].

Evidence on the effects of maternal DBP exposure and SGA (small-for-gestational-age) effects is mixed. Recent reviews have found limited evidence for association with this effect [25, 26], but a cohort study in Massachusetts demonstrated associations between SGA and drinking water trihalomethanes and mutagenic activity levels during the third trimester in the town of maternal residence [92]. In experimental animals, high prenatal maternal exposure to chloroform, BDCM, haloacetonitriles, or haloacetic acids reduced fetal weight [26, 27].

The evidence of the effects of chlorination disinfection by-products in drinking water and birth defects is inconclusive. Some studies have shown an association between neural tube defects and prenatal maternal DBP [25, 26]. The evidence of effects of water disinfection from five previous studies with a cross-sectional study of all Taiwanese births in years 2001–2003 was reviewed [93] and evidence of an effect of exposure to chlorination by-products on the risk of neural tube defects, urinary system defects, and ventricular septal defects was concluded. However, a recent review of a small number of recent studies reported inconsistent results for an association between drinking water chlorination by-products and risk of all congenital anomalies combined and of specific groups of anomalies, and there was little evidence of an exposure–response relationship [94].

Maternal *smoking* during pregnancy increases the risk of pregnancy loss, stillbirth, and infant mortality [28]. There is some strong evidence that exposure to ETS increases the probability of preterm birth [95]. A large birth cohort study in California found borderline statistical association between maternal serum cotinine concentrations during early pregnancy and late fetal deaths (borderline statistical significance), preterm birth, and a significant association with reduced birth weight (significant) [24].

The association between sudden infant death syndrome (SIDS) and maternal smoking has been firmly established. As reported by Wigle et al. [98], a meta-analysis of 39 epidemiological studies and

an expert panel review concluded that there is sufficient evidence of a causal association between sudden infant death syndrome and postnatal ETS exposure, independent of prenatal maternal active smoking [95]. Sudden infant death syndrome was also associated with paternal smoking, even when the mothers did not smoke.

Active *smoking* by mothers has been shown to significantly reduce the rate of fetal growth, and the effect was shown to be dose-dependent. DiFranza and Lew [97] using data from 23 studies, calculated an odds ratio of 1.82 for the association between maternal smoking and low birth weight (<2,500 g). The risk of intrauterine growth retardation caused by maternal smoking appears to increase with maternal age, from twofold for mothers aged 17 years to fivefold for mothers aged 35 years [98]. Although smoking throughout pregnancy is known to affect birth weight, there is some evidence that the final trimester may be particularly important [99]. In some studies, babies of smoking mothers have been reported to be shorter and to have smaller head circumferences at birth than babies of nonsmoking mothers [98].

Studies on *pesticide* exposures and birth defects have found evidence of associations between paternal pesticide exposure and cryptorchidism [99], although other studies have failed to find a similar association. Cryptorchidism was associated also with maternal serum DDT/DDE and hexachlorobenzene levels [100] and with adipose tissue or maternal serum DDE concentrations [101]. In animal studies, trans-placental exposure to the pesticides DDT/DDE vinclozolin, procymidone, or linuron, produce hypospadias, cryptorchidism, and other abnormalities [102]. The epidemiological evidence of the link between pesticide exposure and hypospadias is not conclusive enough to affirm a link between pesticide exposure and this type of birth defect.

Respiratory System

Lung development is a continuous process from embryonic life to adolescence. At birth, about 85% of the human alveoli are present. Alveolar number and lung surface area increase through childhood and begin to level off between 2 and 4 years of age, whereas lung expansion continues up to 8 years of age [5].

Immature (neonatal) differentiating cells of the respiratory tract are more sensitive to injury following exposure to respiratory toxicants than mature cells, and at dose levels that cause no effects in adult cells. Children are usually physically active, and have greater exposure to air pollutants. Because of their higher metabolic rates, they breathe more rapidly and inhale more pollutants per kilogram of body weight than adults. Their narrower airway passages compared to adults can be more easily obstructed in a greater proportion than in adults. Thus, because of the immaturity of the lungs during childhood and the greater exposures relative to body weight in children than in adults, it is inferred that children may be more susceptible to the effects of respiratory toxicants than adults, whose lung growth is complete.

Air pollution has been extensively studied in relation to respiratory health and in general it has been easier to demonstrate the effects of air pollution on exacerbation of certain conditions than on the causing of disease in previously normal individuals. Both outdoor and indoor air pollution have been identified as potential risk factors for both the initiation/induction and the exacerbation of respiratory diseases, especially asthma. Evidence of the effect of air contaminants such as ozone and ETS on lung function has been demonstrated in both animal and human studies. Air pollution is now linked to SIDS [103].

Indoor Air Pollution Indoor air pollution is the most important source of respiratory exposure to toxicants in children, given the fact that children spend up to 90% of their time indoors and the large range of pulmonary irritants that can be found in the home. Exposure to pollutants in the home environment in developed countries has increased with improved insulation and reduced ventilation and the use of chemical detergents and building or furnishing constituents that contain noxious pulmonary irritants [5]. Infants and young children in particular have little control over their exposure to pollutants in the home environment and are vulnerable to the activities of the adults (particularly ETS). Common indoor air pollutants include nitrogen dioxide, formaldehyde and other volatile organic compounds (VOCs), and ETS.

Environmental Tobacco Smoke Environmental tobacco smoke has been extensively studied in relation

to children's respiratory health. The developing fetus can be involuntarily exposed to tobacco during pregnancy through a smoking mother or by the pregnant mother's exposure to environmental tobacco smoke. Exposure may continue throughout childhood if any of the parents smoke. The early exposures may have persistent adverse effects throughout life.

There is evidence of ETS effects on the increased risk of lower respiratory illness rates, especially in the first year of life, increased rates of chronic middle ear effusion in children, impairment of lung function and exacerbation of certain conditions such as asthma. Exposure to environmental tobacco smoke is a risk factor for sudden infant death syndrome. Children exposed to ETS are more likely to suffer from respiratory illness (bronchitis, pneumonia) and to be hospitalized because of the illness than unexposed children [103]. EPA estimates that between 150,000 and 300,000 of bronchitis and pneumonia cases annually in infants and young children up to 18 months of age are attributable to exposure to ETS. Of these, between 7,500 and 15,000 will result in hospitalization. Furthermore, there is evidence that exposure to ETS leads to increased infant mortality from respiratory illness as well as increased morbidity [104].

Exposure to pollutants during critical periods of lung development may have effects that would not be seen if exposure occurred during adulthood [105]. Maternal smoking during pregnancy was related to impaired lung function in newborn infants [106]. Permanent effects of parental smoking on lung function were found in adults of 30–59 years of age [107], indicating that the impairment of lung function persists through life.

The exacerbating effect of ETS on asthma has been known for some time, and many studies evidence this effect. ETS exposure increases the frequency of episodes and severity of symptoms and the rates of hospitalization in asthmatic children. In addition, ETS exposure is a risk factor for new cases of asthma in children who have not previously displayed symptoms (EPA <http://www.epa.gov/smokefree/healtheffects.html>) [69].

Another group of environmental agents that have been studied in relation to respiratory health are the bioaerosols. Bioaerosols include inhaled allergens (house dust mites and other insects, molds, pets, pollens) and bacterial (lipopolysaccharide) and

fungal (glucans) products. Sensitization to one or more common inhalant allergens is consistently associated with childhood asthma especially in developed countries [107]. An expert panel concluded that house-dust mite allergens produce incident (new onset) asthma and that cat, cockroach, and house-dust mite allergens induce episodes in sensitized asthmatics [96]. However, the relationship between exposure to inhaled allergens in early life and the development of asthma or wheeze in childhood is controversial [107]. Contrarily, it has been hypothesized that early exposure to some allergens may be protective of later development of asthma [109], although this hypothesis was discarded in later studies [110].

Outdoor Air Pollution Air pollution, from both vehicular and stationary sources, is associated with an increase in respiratory symptoms, a decrease in lung function, and the exacerbation of asthma symptoms among asthmatic children. Recent studies have evidenced the relationship between traffic air pollution and incident development of asthma. Most of these effects occur at levels within the ambient air quality standards of most countries that have the potential to affect a large population of children. Air pollution effects contribute significantly to children's respiratory morbidity and have a significant economic impact in terms of health expenses. Most studies attribute respiratory symptoms to particulate matter, although the close correlation of particulate matter levels with nitrogen dioxide and sulfur dioxide levels makes the contribution of individual pollutants difficult to determine [5].

There is a consistent body of evidence that outdoor air pollution is associated with increased respiratory symptoms, such as cough, bronchitis, respiratory infections, and upper respiratory tract illness in children [111]. Several studies have demonstrated an increase in respiratory symptoms such as cough and bronchitis associated with increases in PM₁₀. A recent study in Eastern Germany evaluated the association between respiratory symptoms and air pollution levels before and after political reunification. A reduction in air pollution since reunification is associated with reductions in the rates of chronic cough and bronchitis symptoms in a new cohort of children, suggesting a potentially reversible effect of air pollution [112].

A similar dramatic effect was observed with children who moved out of an area to other areas of higher and lower PM₁₀ concentrations. Those children who moved to areas of lower PM₁₀ concentrations showed higher rates of growth in lung function, whereas the opposite effect was observed in children who moved to areas of higher PM₁₀ concentrations than before [113].

The role of air pollution in the exacerbation of asthma is well established, and there is mounting evidence that air pollution, particularly from traffic, is implicated in the pathogenesis of asthma [114]. Because children with asthma have increased airway reactivity, the effects of air pollution on the respiratory system can be more serious for them. Children with asthma have been shown to experience more respiratory symptoms, use extra medication, produce chronic phlegm, and have more bronchitis following exposure to high levels of particulate pollution [115, 116]. Air pollution is also associated with increased school absenteeism due to respiratory illness [117], and increased admissions to hospital emergency department [118].

The role of the different components of air pollution in the development and exacerbation of asthma is difficult to elucidate. Oxidant gases such as nitrogen oxide and ozone have been associated with asthma prevalence [119]. Other studies suggest that traffic pollution, but probably not NO₂ from traffic, is associated with atopy and wheezing [120]. The authors suggested that diesel particles or some component of those particles, such as polycyclic aromatic hydrocarbons, may be the most important etiologic component. A recent review of the short-term effects of PM₁₀ and NO₂ on respiratory health among children with asthma or *asthma*-like symptoms concluded there were clear effects of PM₁₀ on the occurrence of asthma symptom episodes, and to a lesser extent on cough and PEF (pulmonary expiratory function). Results with respect to NO₂ were inconclusive [121].

Of great relevance have been recent studies that have related proximity to heavily trafficked roads with asthma and reduced lung functions as well as other respiratory symptoms such as wheeze and dry cough. Lin et al. [122] found significant odds ratios for living within 200 m of a street with the highest tertile of traffic density and asthma prevalence, and the children with

asthma were more likely to have truck traffic on their street. Although some studies showed no increased risk [123], the weight of evidence suggests that traffic pollution is associated with the risk of developing asthma [114].

Lung function is also affected by air pollution. Acute exposure to ozone, nitrogen dioxide, sulfur dioxide, and particulate matter is known to cause transient reversible decreases in lung function [124, 125]. Some recent studies indicate that long-term exposure to ozone and related co-pollutants (individually and synergistically) is associated with impairment of lung function capacity in children and adolescents [126], although other studies are not consistent. Further, it has been hypothesized that a fraction of the increases in the prevalence of chronic obstructive lung disease in adults who live in more polluted areas could be the result of exposures that occurred during childhood [127].

Cancer

Childhood cancers are relatively rare diseases during the first 2 decades of life, but are the leading cause of death of children in many countries. The most common types of childhood cancer are lymphoid neoplasms (leukemia, lymphoma) and cancers of the central nervous system. Other kinds of childhood tumors include embryonal tumors of the retina, sympathetic nervous system, kidney, and liver; tumors of bone and soft connective tissues; and certain gonadal neoplasms. Carcinomas in epithelial tissues, the most frequent type of cancer in adults, are rare in children. Neoplasms in adults resulting from known or iatrogenic exposure typically have latency periods of 20 years or more. Thus, cancers in children are qualitatively distinct from those in adults.

Although childhood cancer is a rare disease, the occurrence of new cancer cases among children has continued to rise during the last 2 decades. While medical advances have led to a sharp reduction in mortality from childhood cancer, it is still a group of diseases with potentially devastating outcomes, which can be at least partially prevented; so the study of the environmental causes of cancer is essential. A recent report from the European Automated Childhood Cancer Information System (EACCIS) provides evidence of a 1% increase per year in childhood cancers and 1.5%

increase per year in adolescent cancer in Europe for the period 1970–1999. All of the common types of neoplasms showed significant increases, including leukemias, lymphomas, central nervous system tumors, neuroblastomas, soft tissue sarcomas, retinoblastoma, and germ cell, renal, hepatic, and bone tumors in children and carcinomas, lymphomas, soft-tissue sarcomas, and germ cell and CNS tumors in adolescents. There were in addition significant differences between Eastern and Western Europe with regard to leukemias, lymphomas, carcinomas, and central nervous system tumors in children, and carcinomas, lymphomas, leukemias, and soft-tissue tumors in adolescents. Earlier reports of European populations also noted increases in several of these types of childhood cancers. Similarly, in the USA, there was an overall increase of about 25% in childhood cancers between 1975 and 2000. While the reported increases in the 1980s may have reflected improvements in diagnosis and reporting, it is unlikely that the most recently reported trends reflect the same bias. Evidence from both epidemiological and mutagenicity/carcinogenicity studies suggest that environmental toxicants may be involved at least in the causation of some forms of cancers such as acute lymphoblastic leukemia (ALL), which is most common in industrialized countries.

Most cancers result from the interaction between genetic factors and the environment [128]. In this context, environment is defined broadly and includes diet, alcohol, drugs, tobacco, and all other nongenetic factors as well as classic environmental toxins. It has been estimated that about 80–90% of all cancers are attributable to environmental factors acting in conjunction with both genetic and acquired susceptibility. For cancer in adults, different opportunities for environmental exposure have commonly been considered a major (albeit not exclusive) reason for the geographical distribution of cancer. Geographical differences in incidence of most cancers (except lymphomas) are less marked for childhood than for adult cancer, thus suggesting that the fraction of childhood cancers due to environmental factors is probably lower than for adults [128].

The potentially higher susceptibility of the developing fetus and children to the effect of environmental carcinogens is based on several aspects. Chemical carcinogenesis is a multistep process involving genetic and epigenetic changes in susceptible cells that gain

a selective growth advantage and undergo clonal expansion as the result of activation of protooncogenes and/or inactivation of tumor-suppressor genes. The occurrence of these events is modulated by several factors that themselves change in the course of development. Thus, DNA damage may be repaired by the action of DNA repair enzymes, whose presence and activity may change with age. Similarly, changes in metabolism will determine whether carcinogenic metabolites are formed, thus allowing for the initiation of the cancer process.

DNA repair mechanisms play an important role in cancer protection. DNA repair enzymes have been found to be well expressed in embryos and fetuses. In fact, there are a number of DNA repair genes/activities that have higher expression in fetuses or embryos than in adults. Interference with the synthesis of DNA repair enzymes during development may result in increased susceptibility to cancer during childhood or later in life. An example of the important role of DNA repair mechanisms is the p53 gene, which encodes a protein that modulates DNA repair and cell division. Mutations of the p53 genes are involved in at least 50% of all cancers [129]. p53 mutations have been linked with tobacco smoking [130], and it is possible that p53 mutations occur in the offspring of smoking mothers.

The higher rates of cell proliferation during development can contribute to increased likelihood of carcinogenesis. For example, PAHs and aflatoxin B1 (AFB1) produce liver tumors when administered to newborn rodents but not when administered to older animals, presumably because the liver proliferates rapidly in the developing system but more slowly in older animals [131]. Women who were in their teens at the time of atomic bombings had the greatest risk of radiation-induced breast cancer [132]. Some organs, such as the brain, are fully developed in early childhood, whereas others such as the skeletal system do not achieve maturity until adolescence [131]. This may be the reason why osteosarcoma, the most common bone cancer, peaks in late adolescence, a period of rapid bone growth.

The immune system of the newborn is not fully developed until about 6 months of age. Chemicals that affect the immune system, such as halogenated aromatic hydrocarbons that bind to the Ah receptor [133], may modify the host defense mechanism against infection and cancer [134].

The susceptibility of children to cancer may be influenced by the presence/absence of metabolizing enzymes. The role of these enzymes is potentially complex, especially for those that carry out the phase I (usually detoxifying) reactions. Usually, chemical metabolism protects the adult and fetus from carcinogenicity, but the activity of enzymes can also result in bioactivation, such as in the case of bioactivation of benzo[a]pyrene to its carcinogenic form, which has been demonstrated by the observation of the formation of DNA adducts in mice and monkey fetuses, which probably originated in maternal liver [135].

The first step in metabolism is usually oxidation, carried out by cytochrome P450 enzymes. This group of enzymes are named as CYP enzymes and grouped in families according to their genotype and activity. Several groups of these enzymes are involved in the metabolism of exogenous carcinogens, such as the CYP families 1–3, which are thought to be the most relevant to effects of exogenous carcinogens. Other examples are the CYP4B1, which activates aromatic amines, and CYP2E1, which metabolizes nitrosamines and organic solvents that have been implicated in causation of childhood cancers.

Phase II detoxification enzymes have received less attention than CYPs in the perinatal context, though they are also likely to be important. For example, among the glutathione S-transferases (GSTs), the I (P) form is apparently expressed at the highest levels and in many tissues in human fetuses from early in gestation, and levels of total GST activity in all fetal tissues studied was comparable to that in the corresponding adult organ [135]. An increasing body of evidence implicates GST polymorphisms in risk of childhood leukemias. Glucuronidation, a type of Phase II reaction, catalyzed by uridine diphosphate-glucuronosyltransferase (UDG), is a key detoxification step for solubilization of high molecular weight carcinogens such as polycyclic aromatic hydrocarbons, heterocyclic amines, and tobacco-specific nitrosamines. Low expression of UDG in the fetus and neonate, and as a result of polymorphisms in children, could well increase sensitivity to carcinogenesis by some chemicals. Sulfotransferases (SULT) have been proposed to constitute a major enzymatic detoxification system in the fetus [136, 137]. However, these enzymes also catalyze activation of several types of

chemical carcinogens to DNA-damaging forms, including aromatic amines, polycyclic aromatic hydrocarbons, heterocyclic amines, 3-nitrobenzanthrone, and benzylic alcohols. This illustrates how alteration of enzymatic metabolism can greatly influence susceptibility to perinatal carcinogenesis.

While mutagenicity/carcinogenicity studies, as well as the study of enzymatic systems and DNA repair enzymes, provide mounting evidence that the interplay between chemical agents and these biological factors are significant in the development of cancer, epidemiological studies have only been of limited usefulness to confirm this hypothesis. The increasing trends over the last 20–30 years and the geographical distribution of childhood cancers, clearly point to environmental causes. However, few studies have been able to firmly establish the environmental origin of childhood cancers. Cumulative epidemiological evidence pertaining to childhood cancers, including international variation, time trends, and risk factor studies was reviewed and analyzed by Bunin [138]. This review concluded that ionizing radiation and a variety of genetic conditions are thought to explain 5–10% of childhood cancers. There are clear associations between Epstein-Barr virus infection and Burkitts lymphoma in Africa, and between human immunodeficiency virus and Kaposi's sarcoma [138].

Other risk factors have not been conclusively identified. Among the nongenetic causes, the pattern of international variation and associations with surrogates of infection suggested an infectious etiology for acute lymphoblastic leukemia, although no agent has been identified. For brain tumors, cured meats, polyomaviruses, and farm exposures were pointed out as potential causes. Changes in the incidence and characteristics of children with hepatoblastoma as well as risk factor studies suggest a role for an exposure of very low birth weight babies. High birth weight, tea or coffee consumption, and certain paternal occupations have shown some consistency in their association with Wilms' tumor. For most of the other cancers, very few epidemiologic studies have been conducted, so it is not surprising that nongenetic risk factors have not been detected. The most important difference between the cancers for which there are good etiologic clues and those for which there are not may be the number of relevant studies.

This section summarizes the evidence found in the recent literature of the associations between environmental factors and childhood cancer. Attention is given only for those agents for which there is at least limited evidence of an association between environmental agent and disease.

Ionizing Radiation

Evidence of the potential of ionizing radiation to induce childhood cancer comes from events such as Hiroshima and Nagasaki, populations affected by accidents at nuclear plants, as well as investigations on the longer effects of indoor radon and on the consequences of exposure to x-ray for diagnostic or therapeutic reasons.

Japanese children exposed to the atomic bombings of Hiroshima and Nagasaki had increased risks of adult cancers, including leukemia and solid tumors. The highest risk was that of children who were exposed in utero. Radiation-related leukemia started to occur 2–3 years after the bombing, reached its peak within 6–8 years, and has declined steadily since then. For people exposed as adults, the excess risk was lower than that of people exposed as children, but the excess risk appears to have persisted throughout the follow-up period [5].

Thyroid adult carcinomas were also increased in survivors who were children at the time of the bombings at Hiroshima and Nagasaki. The greatest risk for thyroid cancer occurred in individuals who received a radiation dose to the thyroid greater than 1 Sv before age of 10 years. Similarly, breast cancer risk among women survivors to the atomic bombings was highest for the women who were less than 10 years of age or between 10 and 20 years of age, lower for women who were between 20 and 40, and even lower for women exposed after 40 years of age [132].

Prenatal diagnostic X-irradiation has also been linked to increased risk of leukemia in offspring, as has therapeutic, high-dose, ionizing radiation in childhood for other cancers and for various non-neoplastic conditions [139].

The Chernobyl nuclear reactor accident in 1996 caused a significant increase in the incidence of thyroid cancer in children since 1990. Most of the tumors have been observed among individuals who were very young

at the time of the accident [31]. In Belarus, over half of the tumors occurred in people who were less than 6 years old at the time of the accident. In a series of 472 children with thyroid cancer diagnosed up to 1995 in Belarus, only 2% had been conceived after the accident, 9% had been exposed in utero, and 88% were under 15 years of age at the time of diagnosis [140]. The very early age at which the thyroid cancers have begun to be diagnosed is one of the most striking examples of the special sensitivity to a carcinogen other than an infectious agent occurring exclusively during preadult life.

UV Light

Sun exposure is known to be a risk factor for the development of skin cancer later in life. The occurrence of sunburn is an indicator of risk [141]. Numerous studies have assessed the carcinogenic effect of sunburn at different ages and concluded exposure to solar radiation before 10 years of age was a primary contributor to risk of melanoma. The IARC concluded that childhood is an especially vulnerable life stage [142] to the carcinogenic effect of UV radiation. While the use of sunscreen reduces the risk of sunburn, an expert working group convened by the International Agency for Research warned against the risk of relying solely on sunscreens for protection from ultraviolet radiation [143], as the use of sunscreens may lead to an extension of the duration of intentional sun exposure, which can increase the risk of melanoma.

Environmental Tobacco Smoke

Environmental tobacco smoke is an established human carcinogen by the International Agency for Research on Cancer [144]. Over 50 epidemiological studies have reported risk ratios of lung cancer for secondhand smoking in adults. A recent meta-analysis of epidemiological studies of lung cancer and adult exposure to environmental tobacco smoke resulted in risk ratios of 1.22 in women and 1.36 in men from workplace exposure. Other meta-analyses have produced similar results. The evidence of a causal relationship between ETS exposure and cancers in organs other than the lung is inconclusive.

Tobacco smoke contains many carcinogens, such as PAHs and 4-aminobiphenyl, that can cross the placenta

and be transferred to the fetus. The genotoxicity of tobacco smoke to the fetal liver has been tested in an animal study. Sister chromatid exchange in the liver cells of fetal mice was analyzed at the 16th day of gestation after short-term exposure (twice, on the 15th and 16th days of gestation), long-term exposure (starting 4 weeks before mating and stopping on the 16th day of gestation), and prepregnancy exposure (4 weeks before mating). The number of sister chromatid exchanges was significantly increased in all exposure groups, and long-term exposure caused a significantly higher increase than did short-term exposure [145].

Most epidemiological studies on the link between exposure to ETS and childhood cancer have focused mainly on pregnant women. While most of studies of smoking by mothers reveal no effects on childhood cancer, many studies of smoking by fathers have shown a significant association with risk of cancer in their children, and when both mothers and fathers are included in the study, the effect appears to be greater for exposure of fathers than of mothers [135]. This effect has been attributed to germ-cell mutations during spermatogenesis caused by tobacco products [146]. Three reports from the Oxford Survey of Childhood Cancers have suggested that paternal but not maternal cigarette smoking is associated with increased risks for the generality of childhood cancers. Some, however, have produced conflicting findings. A large case control study conducted in the UK (the United Kingdom Childhood Cancer Study) concluded that there was no evidence that paternal smoking is a risk factor for childhood cancer in general [147]. A significant association was reported in this study, however, between hepatoblastoma risks and smoking by both parents relative to neither parent smoking. Sorahan and Lancashire speculated that the importance of both parents smoking in the etiology of hepatoblastoma might arise from the combination of oxidative damage to sperm DNA and damage to the fetal liver from carcinogenic metabolites in the blood of the pregnant mother [148]. Further, in a study where both preconception and postnatal smoking by the fathers was quantified, and few of the mothers smoked, the association with childhood cancer related significantly only to preconceptional paternal smoking levels [149]. Thus, second-hand exposure of the infants to smoke was less

likely. Transplacental effects of sidestream smoke were one possibility. In rats, sidestream smoke constituents received transplacentally caused oxidative DNA damage in fetal tissues [136]. In one epidemiological study, paternal smoking had a larger possible effect when mothers were nonsmokers [137], suggesting protective effects of detoxification enzymes induced in the placenta and in maternal tissues by maternal smoking.

Some studies have examined the relationship between exposure to environmental tobacco smoke during childhood and cancer risk. Sandler and colleagues [150] found that the overall cancer risk was greater for individuals with exposures to environmental tobacco smoke during both childhood and adulthood than for individuals with exposure during only one period. When specific cancer sites or types were considered, leukemia and lymphoma among adults were significantly related to exposure to maternal passive smoke before 10 years of age [150].

Pesticides and Cancer

Pesticides are biologically active molecules that are commonly used to destroy unwanted organisms in agricultural and residential environments. The widespread use of these chemicals has raised concerns over the potential of pesticides to cause childhood cancer.

Although the biochemical mechanisms relating pesticide exposures to childhood cancer have not been fully described, some evidence suggests that pesticides may promote the formation of chromosomal aberrations known to be associated with an increased cancer risk [149]. Studies in adult populations suggest that pesticide exposures may have a direct effect on chromosome structure, and therefore a causal relationship between pesticide exposures and childhood cancer is plausible [152].

Epidemiological studies have reported associations between childhood cancer and either parental or child exposures to pesticides. Research reviews have suggested an increase in the risk of brain cancer, leukemia, non-Hodgkin's lymphoma (NHO), Wilms' tumor, Ewing's sarcoma, and germ cell tumors associated with parental occupational and nonoccupational exposure to pesticides [153, 154]. The exposures observed occurred prior to and during pregnancy, as well as after the childbirth, thus involving different

potential modes of action. Zahm and Ward [154] concluded that at least some childhood cancer could potentially be prevented by reducing or eliminating pesticide exposure, although methodological limitations common to many studies limit the possibility of making conclusions regarding the role of pesticides in the etiology of childhood cancers [153, 154].

Two extensive reviews were conducted after the Zahm and Ward review [154] that attempted to provide conclusions on the basis of cumulative evidence. In a review that evaluated 18 new studies conducted between 1998 and 2004, it was concluded that while collectively all studies seem to suggest an increase in the risk of different cancer types associated with exposure to pesticides, no conclusions could be drawn with respect to cancer types as well as to specific causative factors across studies [153]. Infante-Rivard and Weichenthal [152] reviewed studies conducted between 1999 and 2004 and critically evaluated the evidence on the associations between pesticide exposures and leukemia (12 studies), brain cancer (10 studies), neuroblastoma (4 studies), non-Hodgkin's lymphoma (3 studies), Wilms' tumor (2 studies), and Ewing's sarcoma (1 study), as in the Zahm and Ward review. The authors found recent studies to be consistent with the suggestion of Zahm and Ward of an association between pesticide exposure and childhood leukemia, brain tumors, neuroblastoma, and also non-Hodgkin's lymphoma and Wilms' tumor. Specifically childhood exposure to household insecticides and prenatal exposure to pesticides seemed to pose the greatest risks of leukemia and brain tumors. Exposure-response gradients were observed in some studies on pesticide exposure and risk of leukemia. For neuroblastoma, the authors found that all four recent studies reviewed showed an association between exposure of pesticides and risk of neuroblastoma that supported earlier findings by Zahm and Ward. A recent study also indicates that residential use of pesticides, and herbicides specifically, may increase the risk of neuroblastoma in children [153]. The risk of non-Hodgkin's lymphoma was found to be associated with residential exposure to pesticides, and two of the studies reviewed provided evidence of exposure response gradients in two of the three reviewed studies. With respect to Wilms' tumor, recent studies did not provide additional evidence of an association between insecticide exposures and parental

pesticide exposure before birth and Wilms' tumor that had been indicated in Zahm and Ward's review. The evidence on Ewing's sarcoma was inconclusive.

A systematic review and meta-analysis of studies on childhood leukemia and parental occupational exposure to pesticides concluded that there was not sufficient evidence to affirm an overall association between childhood leukemia and any paternal occupational pesticide exposure among all studies combined or subgroups of studies [155]. An elevated childhood leukemia risk was found in relation to paternal occupational exposure to the broad pesticide classes of insecticides and herbicides, but the authors considered that the small number of studies showing this effect and the lack of exposure–risk relationships did not allow making firm conclusions. However, an association was found between childhood leukemia and prenatal maternal occupational pesticide exposures. The studies also showed associations between childhood leukemia and maternal occupational exposure to insecticides and herbicides. It was concluded that the overall evidence, though limited, warranted exposure prevention measures on the basis of the precautionary principle [152].

Many authors have pointed out the major shortcomings of epidemiological studies on pesticides exposures and childhood cancer [156]. One of the major shortcomings is exposure assessment, as many of the studies use very unspecific measures of exposure, such as “farming” as a measure of “exposures to pesticides.” In most of the studies, the specific pesticide/s, timing and other concomitant exposures to which the population was exposed is not known. The authors recommended directing research efforts to the characterization of pesticide exposures in future research. Further they have remarked that “another 40 epidemiological studies similar to the majority of those conducted thus far will not provide clarity.”

In spite of the limitations of epidemiological research on pesticide exposures and childhood cancers, many researchers find that there is sufficient evidence to be concerned about the potential role of pesticides in childhood cancers, and to recommend a precautionary approach. Infante-Rivard [152] contrasted the overall evidence against the Bradford Hill's causality criteria (strength, consistency, specificity, temporality, biological gradient, plausibility, coherence, experiment and

analogy [157]). The strongest evidence in support of a causal relationship between pesticide exposure and childhood cancer was considered to be the repeated detection of statistically significant increased risks between childhood pesticide exposure and cancer. The authors concluded that there is sufficient evidence to conclude that there is at least some association between pesticide exposure and childhood cancer. In addition, the biological gradients observed in recent studies also suggest that there may be a causal relationship between childhood insecticide exposures and the development of ALL and NHL [158, 159]. Infante-Rivard [152] indicated that the development of childhood cancer probably depends on the presence of many factors, including genetic predisposition, and recommended the use of improved exposure assessments that include separate parental interview, specific pesticide exposure questions, and semiquantitative exposure measures that can be used to confirm exposure information obtained through questionnaires.

Disinfection By-products and Cancer

Disinfection by-products (DBPs), such as trihalomethanes (THMs), are regulated carcinogens in drinking water and have been detected in the blood and breath of swimmers and of nonswimmers at indoor pools. There are a few data on the effects of low doses on humans, particularly infants and children.

Villanueva et al. [160] conducted a pooled analysis of six epidemiological studies and calculated a summary relative risk of bladder cancer equal to 1.18 (95% CI 1.06, 1.32) for exposure above 1 µg/L of trihalomethanes. Boffetta [161] estimated the attributable fraction of bladder cancer on the basis of these figures to be about 10.3%.

Arsenic

Inorganic arsenic is a human carcinogen [162] which causes bladder, skin, and lung cancers in humans. At least in mice, inorganic arsenic is a much more potent carcinogen to the fetus than to adults. The modes of carcinogenic action of inorganic arsenic in rodents and in humans are not yet fully understood, but the possibility exists that inorganic arsenic in drinking water poses a special carcinogenic concern for pregnant women and their unborn infants.

Air Pollution and Childhood Cancer

There have been several studies during the last decade that have linked children's exposure to air pollutants to childhood cancer, leukemia in particular. Infante-Rivard [163] evaluated the results of epidemiological studies conducted between 1998 and 2008 on diverse chemical exposures and childhood leukemia. The review included nine case-control studies and four ecological studies. The latter showed an association between incidence rates of childhood leukemia and levels of air pollution in the area of residence of the population at the time of diagnosis. The lack of information about the place of residence of subjects prior to diagnosis (including during the preconception and gestation period) made it difficult to conclude an association between the exposure levels and risk of childhood leukemia. The review concluded that the weight of evidence from both case-control and ecological studies indicated no increased risk for childhood leukemia associated with exposure to traffic-related residential air pollution. The same conclusion was reached in another review conducted by Raaschou-Nielsen [164].

Endocrine Disruptors and Cancer

Increases in the incidence of cancer in certain parts of the world are often cited as evidence that widespread exposure to endocrine disruptors has adverse effects on human health. Of particular concern are the observed increased incidences of cancer at hormonally sensitive sites, such as breast, uterus, prostate, and testis in Europe and North America. These increases cannot be explained only on the basis of improved diagnostic techniques, and it has been argued that these trends coincide roughly with the increasing use and release of industrial chemicals into the environment. Concerns are also based on plausible mechanisms of action because both human and experimental animal studies show that these cancers are modulated hormonally.

Many epidemiological and experimental animal studies have attempted to evaluate the link between exposure to endocrine disruptors and increased risk of breast cancer. A direct association between these chemicals and increased risk of breast cancer has not been established. However, some researchers claim that the time of life when exposure takes place

(e.g., prenatal, neonatal, childhood, adolescence) may have a major influence on the appearance of this and other health effects. The development of the mammary gland occurs in multiple stages. Fetal development of the mammary gland rudiment is governed by tissue interactions in both males and females. In females, the pubertal period drives ductal morphogenesis, and pregnancy results in massive differentiation of the mammary gland. Thus, the perinatal period and the period between age at menarche and age at first full-term pregnancy may be particularly important for breast tumor development and latency [165]. Young girls exposed to carcinogenic agents during puberty may be at high risk of future breast cancer due to susceptibility of rapidly growing breast tissue mediated by hormonal changes during this time. This claim is supported by data from atomic bomb survivors, where an increased risk of breast cancer was found in women exposed before 20 years of age [132]. Similarly, an elevated risk was found for women irradiated during childhood for medical reasons [166].

DES was extensively prescribed in developed countries from the late 1940s through the 1970s to women with high-risk pregnancies to prevent miscarriages and other complications of pregnancy. In the early 1970s, a rare form of female reproductive tract cancer, clear cell adenocarcinoma began to be detected among women whose mothers had taken DES during pregnancy [167]. Although clear cell adenocarcinoma occurs in only 0.1% of women who were exposed to DES in utero, this represents a 40-fold increased risk in comparison with the nonexposed population. In contrast, men who were exposed to DES in utero do not have a clear increased risk of any cancer, although a statistically nonsignificant threefold increased risk of testicular cancer has been reported [168]. While there is no epidemiological evidence for a link between exposure to estrogenic or anti-androgenic compounds and testicular cancer, some authors have hypothesized that the similar increases in the incidence of testicular cancer and in the incidence of cryptorchidism and hypospadias in similar geographical areas, suggest a common cause of similar environmental origin in both health effects [85].

Some studies have shown that women taking DES during pregnancy to prevent miscarriage have been shown to have a slightly increased risk of developing

breast cancer 30 years after taking the drug [169]. Data on the risks of breast cancer in daughters of DES-exposed women are not yet available.

Future Directions

During the last 2 decades, there has been an exponential increase of scientific literature about the susceptibility of children to the effects of environmental exposures. Evidence is accumulating that toxic chemicals are responsible for at least some of the changing patterns of disease. Well-studied examples include adenocarcinoma of the vagina in girls exposed prenatally to DES; asthma and pneumonia caused by smoke and particulate air pollutants; neurodevelopmental toxicity in infants exposed to lead, PCBs, methylmercury; and small head circumference at birth in infants exposed in utero to organophosphate pesticides.

Past discoveries of etiologic associations between toxic environmental exposures and diseases in children have led to successful programs of exposure control and disease prevention. Examples include reductions in the use of alcohol and tobacco during pregnancy, minimization during pregnancy of diagnostic x-rays, and removal of lead from gasoline. Sadly, though, the interval between initial recognition or suspicion of effects and their eventual control has been typically been far too long. Early warnings have frequently been ignored. The price of delayed action has been widespread increases in the incidence of certain diseases such as asthma and cancer.

When evaluating the health and social impact of environmental threats on children, it is necessary to take into account, not only their effects on childhood health, but also the long-term potential health effects throughout the lifelong span of the individual. It has been hypothesized that early exposure to environmental toxicants could affect the brain later in life. Consensus among scientists is based on experimental studies on associations between early life exposures to pesticides and Parkinson's disease, as well as on epidemiologic studies of the toxic and apparently irreversible effects on the developing brain of in utero exposure to lead, methylmercury, and polychlorinated biphenyls. A mechanistic hypothesis proposed that early exposure to neurotoxic chemicals reduce the number of neurons in critical areas of the brain such as the substantia nigra

to levels below those needed to sustain function in the face of neuronal attrition associated with advancing age. In addition, as some researchers have pointed out, the effects at population level would add a substantial burden to society, in terms of health, economic, and human costs.

The difficulties in conducting environmental health research have been pointed out by many authors during the last 3 decades. Human populations are exposed to hundreds of chemicals through air, water, and food, under many different exposure situations. Environmental exposures are usually low, and often occur concomitantly with occupational exposures, smoking or naturally occurring agents. And when an association between with an exposure and a disease is found, such as for example, air pollution and asthma, a long path lies ahead to determine the specific causative agent of the disease, and the population and individual susceptibility factors that makes some individuals and not others become ill. The potential for interaction of different pollutant exposures makes the question with respect to the role of environmental exposures impact on health even more difficult to answer. The need for taking into account multiple and cumulative exposures has been clearly affirmed by the EPA and other government and research bodies.

Slowly, evidence is found to confirm or discard certain effects, but it seems that progress is too slow to catch up to the increases in incidence of certain diseases, such as asthma and cancer. It is possible that the statement of Olshan and Daniels [170] with respect to pesticide research, that "another 40 epidemiological studies similar to the majority of those conducted thus far will not provide clarity," should be applied to other groups of chemicals, such as disinfection by-products, or endocrine disruptors, and that we need to think of a more effective way of examining evidence. Many authors have pointed out the need of better exposure assessment methods. In addition to continuing toxicological research, which is essential to elucidate the mechanisms of action, epidemiological research should probably incorporate exposure assessment methods that could enable detection of at least the specific exposures and exposure levels. Exposure assessment can be improved in many ways. In the extreme, if the body burden of a list of suspected carcinogens in the blood or adipose tissue of each

child diagnosed with cancer were determined, more specific associations between disease and toxic agent may be determined. The path to confirm and rule out some proposed environmental causes would probably be at least somewhat shorter. While this approach may be costly, it may be cost-efficient in the long run. There have been some efforts to integrate an environmental health perspective in the medical practice. These efforts should be further developed and supported.

Bibliography

Primary Literature

1. Tamburlini G, von Ehrenstein OS, Bertollini R (eds) (2002) Children's health and environment: A review of evidence. A joint report from the European Environment Agency and the WHO Regional Office for Europe. Environmental issue report No. 29. Office for Official Publications of the European Communities, Luxembourg
2. NRC (1993) Pesticides in the diets of infants and children. US National Research Council, National Academy Press, Washington, DC
3. ILSI (1992) Similarities and differences between children and adults: implications for risk assessment. International Life Sciences Institute Press, Washington, DC
4. ILSI (2003) Final report: workshop to develop a framework for assessing risks to children from exposure to environmental agents. International Life Sciences Institute Press, Washington, DC
5. WHO (2006) Principles for evaluating health risks in children. World Health Organization, Geneva
6. Moya J, Bearer CF, Etzel RA (2004) Children's behavior and physiology and how it affects exposure to environmental contaminants. *Pediatrics* 113(4 Suppl):996–1006
7. Landrigan PJ (2004) Children as a vulnerable population. *Int J Occup Med Environ Health* 17(1):175–177
8. Calabrese EJ, Stanek EJ, James RC, Roberts SM (1997) Soil ingestion: a concern for acute toxicity in children. *Environ Health Perspect* 105(12):1354–1358
9. Simcox NJ, Fenske RA, Wolz SA, Lee IC, Kalman DA (1995) Pesticides in household dust and soil: exposure pathways for children of agricultural families. *Environ Health Perspect* 103(12):1126–1134
10. Clewell HJ, Teeguarden J, McDonald T, Sarangapani R, Lawrence G, Covington T et al (2002) Review and evaluation of the potential impact of age- and gender-specific pharmacokinetic differences on tissue dosimetry. *Crit Rev Toxicol* 32(5):329–389
11. Anderson LM, Diwan BA, Fear NT, Roman E (2000) Critical windows of exposure for children's health: cancer in human epidemiological studies and neoplasms in experimental animal models. *Environ Health Perspect* 108(Suppl 3): 573–594
12. Loeffler IK, Peterson RE (1999) Interactive effects of TCDD and p, p'-DDE on male reproductive tract development in in utero and lactationally exposed rats. *Toxicol Appl Pharmacol* 154(1):28–39
13. Weinberg CR, Wilcox AJ, Baird DD (1989) Reduced fecundability in women with prenatal exposure to cigarette smoking. *Am J Epidemiol* 129(5):1072–1078
14. Wilcox AJ, Baird DD, Weinberg CR, Hornsby PP, Herbst AL (1995) Fertility in men exposed prenatally to diethylstilbestrol. *N Engl J Med* 332(21):1411–1416
15. Buckley JD, Robison LL, Swotinsky R, Garabrant DH, LeBeau M, Manchester P et al (1989) Occupational exposures of parents of children with acute nonlymphocytic leukemia: a report from the Children's Cancer Study Group. *Cancer Res* 49(14):4030–4037
16. Nicol CJ, Harrison ML, Laposa RR, Gimelshtein IL, Wells PG (1995) A teratologic suppressor role for p53 in benzo[a]pyrene-treated transgenic p53-deficient mice. *Nat Genet* 10(2):181–187
17. Mullikin-Kilpatrick D, Mehta ND, Hildebrandt JD, Treisman SN (1995) Gi is involved in ethanol inhibition of L-type calcium channels in undifferentiated but not differentiated PC-12 cells. *Mol Pharmacol* 47(5):997–1005
18. Murante FG, Gasiewicz TA (2000) Hemopoietic progenitor cells are sensitive targets of 2, 3, 7, 8-tetrachlorodibenzo-p-dioxin in C57BL/6J mice. *Toxicol Sci* 54(2):374–383
19. Amsen D, Kruisbeek AM (1998) Thymocyte selection: not by TCR alone. *Immunol Rev* 165:209–229
20. Rodier PM (2004) Environmental causes of central nervous system maldevelopment. *Pediatrics* 113(4 Suppl):1076–1083
21. Sharpe CR, Franco EL (1995) Etiology of Wilms' tumor. *Epidemiol Rev* 17(2):415–432
22. Bunney BG, Potkin SG, Bunney WE Jr (1995) New morphological and neuropathological findings in schizophrenia: a neurodevelopmental perspective. *Clin Neurosci* 3(2):81–88
23. Rodier PM (1995) Developing brain as a target of toxicity. *Environ Health Perspect* 103(Suppl 6):73–76
24. Kharrazi M, DeLorenze GN, Kaufman FL, Eskenazi B, Bernert JT Jr, Graham S et al (2004) Environmental tobacco smoke and pregnancy outcome. *Epidemiology* 15(6):660–670
25. Bove F, Shim Y, Zeitz P (2002) Drinking water contaminants and adverse pregnancy outcomes: a review. *Environ Health Perspect* 110(Suppl 1):61–74
26. Graves CG, Matanoski GM, Tardiff RG (2001) Weight of evidence for an association between adverse reproductive and developmental effects and exposure to disinfection by-products: a critical review. *Regul Toxicol Pharmacol* 34(2):103–124
27. Nieuwenhuijsen MJ, Toledano MB, Eaton NE, Fawell J, Elliott P (2000) Chlorination disinfection byproducts in water and their association with adverse reproductive outcomes: a review. *Occup Environ Med* 57(2):73–85
28. Platt RW, Joseph KS, Ananth CV, Grondines J, Abrahamowicz M, Kramer MS (2004) A proportional hazards model with time-dependent covariates and time-varying effects for analysis of fetal and infant death. *Am J Epidemiol* 160(3):199–206

29. Paus T, Zijdenbos A, Worsley K, Collins DL, Blumenthal J, Giedd JN et al (1999) Structural maturation of neural pathways in children and adolescents: in vivo study. *Science* 283(5409):1908–1911
30. Dietert RR, Lee JE, Bunn TL (2002) Developmental immunotoxicology: emerging issues. *Hum Exp Toxicol* 21(9–10):479–485
31. International Agency for Research on Cancer (2001) IARC monographs on the evaluation of carcinogenic risks to humans: ionizing radiation, Part 2: some internally deposited radionuclides, vol 78. International Agency for Research on Cancer, Lyon, France
32. Altschuler K (2003) Critical periods in development. EPA . Paper Series on Children's Health and the Environment
33. Mahaffey KR (1991) Biokinetics of lead during pregnancy. *Fundam Appl Toxicol* 16(1):15–16
34. American Academy of Pediatrics Committee on Environmental Health (2003) Pediatric environmental health, 2nd edn. American Academy of Pediatrics, Elk Grove Village
35. Perera FP, Jedrychowski W, Rauh V, Whyatt RM (1999) Molecular epidemiologic research on the effects of environmental pollutants on the fetus. *Environ Health Perspect* 107(Suppl 3):451–460
36. Rice D, Barone S Jr (2000) Critical periods of vulnerability for the developing nervous system: evidence from humans and animal models. *Environ Health Perspect* 108(Suppl 3):511–533
37. Casey BJ, Tottenham N, Liston C, Durston S (2005) Imaging the developing brain: what have we learned about cognitive development? *Trends Cogn Sci* 9(3):104–110
38. Cory-Slechta DA, Thiruchelvam M, Barlow BK, Richfield EK (2005) Developmental pesticide models of the Parkinson disease phenotype. *Environ Health Perspect* 113(9):1263–1270
39. Landrigan PJ, Whitworth RH, Baloh RW, Staehling NW, Barthel WF, Rosenblum BF (1975) Neuropsychological dysfunction in children with chronic low-level lead absorption. *Lancet* 1(7909):708–712
40. Needleman HL (1979) Lead levels and children's psychologic performance. *N Engl J Med* 301(3):1363
41. Bellinger DC (2004) Lead. *Pediatrics* 113(4 Suppl):1016–1022
42. Lanphear BP, Hornung R, Khoury J, Yolton K, Baghurst P, Bellinger DC et al (2005) Low-level environmental lead exposure and children's intellectual function: an international pooled analysis. *Environ Health Perspect* 113(7):894–899
43. Rothenberg SJ, Poblano A, Schnaas L (2000) Brainstem auditory evoked response at five years and prenatal and postnatal blood lead. *Neurotoxicol Teratol* 22(4):503–510
44. Bellinger D, Hu H, Titlebaum L, Needleman HL (1994) Attentional correlates of dentin and bone lead levels in adolescents. *Arch Environ Health* 49(2):98–105
45. Dietrich KN, Ris MD, Succop PA, Berger OG, Bornschein RL (2001) Early exposure to lead and juvenile delinquency. *Neurotoxicol Teratol* 23(6):511–518
46. Grandjean P, Weihe P, White RF, Debes F, Araki S, Yokoyama K et al (1997) Cognitive deficit in 7-year-old children with prenatal exposure to methylmercury. *Neurotoxicol Teratol* 19(6):417–428
47. Grandjean P, White RF, Nielsen A, Cleary D, Oliveira Santos EC (1999) Methylmercury neurotoxicity in Amazonian children downstream from gold mining. *Environ Health Perspect* 107(7):587–591
48. Davidson PW, Myers GJ, Weiss B (2004) Mercury exposure and child development outcomes. *Pediatrics* 113(4 Suppl):1023–1029
49. Murata K, Weihe P, Budtz-Jorgensen E, Jorgensen PJ, Grandjean P (2004) Delayed brainstem auditory evoked potential latencies in 14-year-old children exposed to methylmercury. *J Pediatr* 144(2):177–183
50. Newland MC, Paletz EM (2000) Animal studies of methylmercury and PCBs: what do they tell us about expected effects in humans? *Neurotoxicology* 21(6):1003–1027
51. National Academy of Sciences CotTEM (2000) Toxicological effects of methylmercury. National Academy Press, Washington, DC
52. Brouwer A, Longnecker MP, Birnbaum LS, Coglian J, Kostyniak P, Moore J et al (1999) Characterization of potential endocrine-related health effects at low-dose levels of exposure to PCBs. *Environ Health Perspect* 107(Suppl 4):639–649
53. Rogan WJ, Gladen BC, Hung KL, Koong SL, Shih LY, Taylor JS et al (1988) Congenital poisoning by polychlorinated biphenyls and their contaminants in Taiwan. *Science* 241(4863):334–336
54. Guo YL, Lambert GH, Hsu CC, Hsu MM (2004) Yucheng: health effects of prenatal exposure to polychlorinated biphenyls and dibenzofurans. *Int Arch Occup Environ Health* 77(3):153–158
55. Chen YC, Yu ML, Rogan WJ, Gladen BC, Hsu CC (1994) A 6-year follow-up of behavior and activity disorders in the Taiwan Yu-Cheng children. *Am J Public Health* 84(3):415–421
56. Ribas-Fito N, Sala M, Kogevinas M, Sunyer J (2001) Polychlorinated biphenyls (PCBs) and neurological development in children: a systematic review. *J Epidemiol Community Health* 55(8):537–546
57. Grandjean P, Weihe P, Burse VW, Needham LL, Storr-Hansen E, Heinzow B et al (2001) Neurobehavioral deficits associated with PCB in 7-year-old children prenatally exposed to seafood neurotoxins. *Neurotoxicol Teratol* 23(4):305–317
58. Guillette EA, Meza MM, Aquilar MG, Soto AD, Garcia IE (1998) An anthropological approach to the evaluation of preschool children exposed to pesticides in Mexico. *Environ Health Perspect* 106(6):347–353
59. Moser VC, Chanda SM, Mortensen SR, Padilla S (1998) Age- and gender-related differences in sensitivity to chlorpyrifos in the rat reflect developmental profiles of esterase activities. *Toxicol Sci* 46(2):211–222
60. Slotkin TA, Tate CA, Cousins MM, Seidler FJ (2002) Functional alterations in CNS catecholamine systems in adolescence and adulthood after neonatal chlorpyrifos exposure. *Brain Res Dev Brain Res* 133(2):163–173
61. EPA (1997) Background document on cholinesterases. Attachment 4E presented to Environmental Protection Agency Scientific Advisory Panel Meeting of June 3–4, 1997. US EPA, Arlington

62. Weiss B, Amler S, Amler WR (2004) Pesticides. *Pediatrics* 113:1030–1036
63. Damstra T (2002) Potential effects of certain persistent organic pollutants and endocrine disrupting chemicals on the health of children. *J Toxicol Clin Toxicol* 40(4):457–465
64. Ho SM, Tang WY, Belmonte DF, Prins GS (2006) Developmental exposure to estradiol and bisphenol A increases susceptibility to prostate carcinogenesis and epigenetically regulates phosphodiesterase type 4 variant 4. *Cancer Res* 66(11):5624–5632
65. Newbold RR, Hanson RB, Jefferson WN, Bullock BC, Haseman J, McLachlan JA (2000) Proliferative lesions and reproductive tract tumors in male descendants of mice exposed developmentally to diethylstilbestrol. *Carcinogenesis* 21(7):1355–1363
66. Anway MD, Cupp AS, Uzumcu M, Skinner MK (2005) Epigenetic transgenerational actions of endocrine disruptors and male fertility. *Science* 308(5727):1466–1469
67. Goldey ES, Crofton KM (1998) Thyroxine replacement attenuates hypothyroxinemia, hearing loss, and motor deficits following developmental exposure to Aroclor 1254 in rats. *Toxicol Sci* 45(1):94–105
68. Brechner RJ, Parkhurst GD, Humble WO, Brown MB, Herman WH (2000) Ammonium perchlorate contamination of Colorado River drinking water is associated with abnormal thyroid function in newborns in Arizona. *J Occup Environ Med* 42(8):777–782
69. Kelsh MA, Buffler PA, Daaboul JJ, Rutherford GW, Lau EC, Barnard JC et al (2003) Primary congenital hypothyroidism, newborn thyroid function, and environmental perchlorate exposure among residents of a Southern California community. *J Occup Environ Med* 45(10):1116–1127
70. Lau C, Rogers JM (2004) Embryonic and fetal programming of physiological disorders in adulthood. *Birth Defects Res C Embryo Today* 72(4):300–312
71. Sadler TW (2000) Susceptible periods during embryogenesis of the heart and endocrine glands. *Environ Health Perspect* 108(Suppl 3):555–561
72. Ravelli AC, van der Meulen JH, Michels RP, Osmond C, Barker DJ, Hales CN et al (1998) Glucose tolerance in adults after prenatal exposure to famine. *Lancet* 351(9097):173–177
73. Bo S, Cavallo-Perin P, Scaglione L, Ciccone G, Pagano G (2000) Low birthweight and metabolic abnormalities in twins with increased susceptibility to Type 2 diabetes mellitus. *Diabet Med* 17(5):365–370
74. Pryor JL, Hughes C, Foster W, Hales BF, Robaire B (2000) Critical windows of exposure for children's health: the reproductive system in animals and humans. *Environ Health Perspect* 108(Suppl 3):491–503
75. Stillman RJ (1982) In utero exposure to diethylstilbestrol: adverse effects on the reproductive tract and reproductive performance and male and female offspring. *Am J Obstet Gynecol* 142(7):905–921
76. Parente RC, Faerstein E, Celeste RK, Werneck GL (2008) The relationship between smoking and age at the menopause: a systematic review. *Maturitas* 61(4):287–298
77. Jensen TK, Joffe M, Scheike T, Skytthe A, Gaist D, Christensen K (2005) Time trends in waiting time to pregnancy among Danish twins. *Hum Reprod* 20(4):955–964
78. Shea KM (2003) Pediatric exposure and potential toxicity of phthalate plasticizers. *Pediatrics* 111(6 Pt 1):1467–1474
79. Weisbach V, Koch HM, Angerer J, Eckstein R (2006) Di(2-ethylhexyl)phthalate exposure of apheresis donors is procedure-related. *Transfusion* 46(8):1457–1458
80. Skakkebaek NE, Jorgensen N, Main KM, Rajpert-De Meyts E, Leffers H, Andersson AM et al (2006) Is human fecundity declining? *Int J Androl* 29(1):2–11
81. Lottrup G, Andersson AM, Leffers H, Mortensen GK, Toppari J, Skakkebaek NE et al (2006) Possible impact of phthalates on infant reproductive health. *Int J Androl* 29(1):172–180
82. Mylchreest E, Wallace DG, Cattley RC, Foster PM (2000) Dose-dependent alterations in androgen-regulated male reproductive development in rats exposed to Di(n-butyl) phthalate during late gestation. *Toxicol Sci* 55(1):143–151
83. Hoei-Hansen CE, Holm M, Rajpert-De Meyts E, Skakkebaek NE (2003) Histological evidence of testicular dysgenesis in contralateral biopsies from 218 patients with testicular germ cell cancer. *J Pathol* 200(3):370–374
84. Skakkebaek NE, Holm M, Hoei-Hansen C, Jorgensen N, Rajpert-De Meyts E (2003) Association between testicular dysgenesis syndrome (TDS) and testicular neoplasia: evidence from 20 adult patients with signs of maldevelopment of the testis. *APMIS* 111(1):1–9
85. Fisher JS, Macpherson S, Marchetti N, Sharpe RM (2003) Human 'testicular dysgenesis syndrome': a possible model using in-utero exposure of the rat to dibutyl phthalate. *Hum Reprod* 18(7):1383–1394
86. Weinhold B (2009) Environmental factors in birth defects: what we need to know. *Environ Health Perspect* 117:A440–A447
87. Lindbohm ML, Taskinen H, Sallmen M, Hemminki K (1990) Spontaneous abortions among women exposed to organic solvents. *Am J Ind Med* 17(4):449–463
88. Lipscomb JA, Fenster L, Wrensch M, Shusterman D, Swan S (1991) Pregnancy outcomes in women potentially exposed to occupational solvents and women working in the electronics industry. *J Occup Med* 33(5):597–604
89. Nieuwenhuijsen MJ, Toledano MB, Elliott P (2000) Uptake of chlorination disinfection by-products: a review and a discussion of its implications for exposure assessment in epidemiological studies. *J Expo Anal Environ Epidemiol* 10(6 Pt 1):586–599
90. Dodds L, King W, Allen AC, Armson BA, Fell DB, Nimrod C (2004) Trihalomethanes in public water supplies and risk of stillbirth. *Epidemiology* 15(2):179–186
91. Toledano MB, Nieuwenhuijsen MJ, Best N, Whitaker H, Hambly P, de Hoogh C et al (2005) Relation of trihalomethane concentrations in public water supplies to stillbirth and birth weight in three water regions in England. *Environ Health Perspect* 113(2):225–232

92. Wright JM, Schwartz J, Dockery DW (2004) The effect of disinfection by-products and mutagenic activity on birth weight and gestational duration. *Environ Health Perspect* 112(8):920–925
93. Hwang BF, Jaakkola JJ (2003) Water chlorination and birth defects: a systematic review and meta-analysis. *Arch Environ Health* 58(2):83–91
94. Nieuwenhuijsen MJ, Martinez D, Grellier J, Bennett J, Best N, Iszatt N et al (2009) Chlorination disinfection by-products in drinking water and congenital anomalies: review and meta-analyses. *Environ Health Perspect* 117(10):1486–1493
95. California Environmental Protection Agency (2010) Proposed identification of environmental tobacco smoke as a toxic air contaminant. California Environmental Protection Agency, Sacramento
96. Wigle DT, Arbuckle TE, Walker M, Wade MG, Liu S, Krewski D (2007) Environmental hazards: evidence for effects on child health. *J Toxicol Environ Health B Crit Rev* 10(1–2):3–39
97. DiFranza JR, Lew RA (1995) Effect of maternal cigarette smoking on pregnancy complications and sudden infant death syndrome. *J Fam Pract* 40(4):385–394
98. Charlton A (1994) Children and passive smoking: a review. *J Fam Pract* 38(3):267–277
99. Wang J, Wang B (2002) Study on risk factors of cryptorchidism. *Zhonghua Liu Xing Bing Xue Za Zhi* 23(3):190–193
100. Waliszewski SM, Infanzon RM, Arroyo SG, Pietrini RV, Carvajal O, Trujillo P et al (2005) Persistent organochlorine pesticides levels in blood serum lipids in women bearing babies with undescended testis. *Bull Environ Contam Toxicol* 75(5):952–959
101. Longnecker MP, Klebanoff MA, Dunson DB, Guo X, Chen Z, Zhou H et al (2005) Maternal serum level of the DDT metabolite DDE in relation to fetal loss in previous pregnancies. *Environ Res* 97(2):127–133
102. Gray LE, Ostby J, Furr J, Wolf CJ, Lambright C, Parks L et al (2001) Effects of environmental antiandrogens on reproductive development in experimental animals. *Hum Reprod Update* 7(3):248–264
103. Woodruff TJ, Zeise L, Axelrad DA, Guyton KZ, Janssen S, Miller M et al (2008) Meeting report: moving upstream-evaluating adverse upstream end points for improved risk assessment and decision-making. *Environ Health Perspect* 116(11):1568–1575
104. DiFranza JR, Aligne CA, Weitzman M (2004) Prenatal and postnatal environmental tobacco smoke exposure and children's health. *Pediatrics* 113(4 Suppl):1007–1015
105. Dietert RR, Etzel RA, Chen D, Halonen M, Holladay SD, Jarabek AM et al (2000) Workshop to identify critical windows of exposure for children's health: immune and respiratory systems work group summary. *Environ Health Perspect* 108(Suppl 3):483–490
106. Stick SM, Burton PR, Gurrin L, Sly PD, LeSouef PN (1996) Effects of maternal smoking during pregnancy and a family history of asthma on respiratory function in newborn infants. *Lancet* 348(9034):1060–1064
107. Upton MN, Watt GC, Davey SG, McConnachie A, Hart CL (1998) Permanent effects of maternal smoking on offsprings' lung function. *Lancet* 352(9126):453
108. Sporik R, Platts-Mills TA (2001) Allergen exposure and the development of asthma. *Thorax* 56(Suppl 2):ii58–ii63
109. Hesselmar B, Aberg N, Aberg B, Eriksson B, Bjorksten B (1999) Does early exposure to cat or dog protect against later allergy development? *Clin Exp Allergy* 29(5):611–617
110. Merchant JA, Naleway AL, Svendsen ER, Kelly KM, Burmeister LF, Stromquist AM et al (2005) Asthma and farm exposures in a cohort of rural Iowa children. *Environ Health Perspect* 113(3):350–356
111. Schwela D (2000) Air pollution and health in urban areas. *Rev Environ Health* 15(1–2):13–42
112. Heinrich J (2003) Nonallergic respiratory morbidity improved along with a decline of traditional air pollution levels: a review. *Eur Respir J Suppl* 40:64s–69s
113. Avol EL, Gauderman WJ, Tan SM, London SJ, Peters JM (2001) Respiratory effects of relocating to areas of differing air pollution level. *Am J Respir Crit Care Med* 164(11):2067–2072
114. Schwartz J (2004) Air pollution and children's health. *Pediatrics* 113(4 Suppl):1037–1043
115. Ostro B, Lipsett M, Mann J, Braxton-Owens H, White M (2001) Air pollution and exacerbation of asthma in African-American children in Los Angeles. *Epidemiology* 12(2):200–208
116. McConnell R, Berhane K, Gilliland F, Molitor J, Thomas D, Lurmann F et al (2003) Prospective study of air pollution and bronchitic symptoms in children with asthma. *Am J Respir Crit Care Med* 168(7):790–797
117. Gilliland FD, Berhane K, Rappaport EB, Thomas DC, Avol E, Gauderman WJ et al (2001) The effects of ambient air pollution on school absenteeism due to respiratory illnesses. *Epidemiology* 12(1):43–54
118. Tolbert PE, Mulholland JA, MacIntosh DL, Xu F, Daniels D, Devine OJ et al (2000) Air quality and pediatric emergency room visits for asthma in Atlanta, Georgia, USA. *Am J Epidemiol* 151(8):798–810
119. Studnicka M, Hackl E, Pischinger J, Fangmeyer C, Haschke N, Kuhr J et al (1997) Traffic-related NO₂ and the prevalence of asthma and respiratory symptoms in seven year olds. *Eur Respir J* 10(10):2275–2278
120. Kramer U, Koch T, Ranft U, Ring J, Behrendt H (2000) Traffic-related air pollution is associated with atopy in children living in urban areas. *Epidemiology* 11(1):64–70
121. Weinmayr G, Genuneit J, Nagel G, Bjorksten B, van Hage M, Priftanji A et al (2010) International variations in associations of allergic markers and diseases in children: ISAAC phase two. *Allergy* 65(6):766–775
122. Lin S, Munsie JP, Hwang SA, Fitzgerald E, Cayo MR (2002) Childhood asthma hospitalization and residential exposure to state route traffic. *Environ Res* 88(2):73–81

123. English P, Neutra R, Scalf R, Sullivan M, Waller L, Zhu L (1999) Examining associations between childhood asthma and traffic flow using a geographic information system. *Environ Health Perspect* 107(9):761–767
124. Vedal S, Petkau J, White R, Blair J (1998) Acute effects of ambient inhalable particles in asthmatic and non-asthmatic children. *Am J Respir Crit Care Med* 157(4 Pt 1): 1034–1043
125. Pekkanen J, Pearce N (1999) Defining asthma in epidemiological studies. *Eur Respir J* 14(4):951–957
126. Gauderman WJ, Gilliland GF, Vora H, Avol E, Stram D, McConnell R et al (2002) Association between air pollution and lung function growth in southern California children: results from a second cohort. *Am J Respir Crit Care Med* 166(1):76–84
127. American Academy of Pediatrics Committee on Environmental Health (2003) Air pollutants, outdoor. In: Etzel R (ed) *Pediatric environmental health*. American Academy of Pediatrics, Elk Grove Village, pp 69–86
128. Perera FP (1997) Environment and cancer: who are susceptible? *Science* 278(5340):1068–1073
129. Bartsch H, Hietanen E (1996) The role of individual susceptibility in cancer burden related to environmental exposure. *Environ Health Perspect* 104(Suppl 3):569–577
130. Husgafvel-Pursiainen K (2004) Genotoxicity of environmental tobacco smoke: a review. *Mutat Res* 567(2–3):427–445
131. National Research Council (1982) *Diet, nutrition and cancer*. National Academy Press, Washington, DC
132. Tokunaga M, Land CE, Tokuoka S, Nishimori I, Soda M, Akiba S (1994) Incidence of female breast cancer among atomic bomb survivors, 1950–1985. *Radiat Res* 138(2):209–223
133. Schiestl RH, Aubrecht J, Yap WY, Kandikonda S, Sidhom S (1997) Polychlorinated biphenyls and 2, 3, 7, 8-tetrachlorodibenzo-p-dioxin induce intrachromosomal recombination in vitro and in vivo. *Cancer Res* 57(19):4378–4383
134. Trizio D, Basketter DA, Botham PA, Graepel PH, Lambre C, Magda SJ et al (1988) Identification of immunotoxic effects of chemicals and assessment of their relevance to man. *Food Chem Toxicol* 26(6):527–539
135. Anderson LM (2006) Environmental genotoxicants/carcinogens and childhood cancer: bridgeable gaps in scientific knowledge. *Mutat Res* 608(2):136–156
136. Maciag A, Bialkowska A, Espiritu I, Powell D, Alvord WG, Kasprzak KS et al (2003) Gestation stage-specific oxidative deoxyribonucleic acid damage from sidestream smoke in pregnant rats and their fetuses. *Arch Environ Health* 58(4):238–244
137. Schuz J, Kaletsch U, Kaatsch P, Meinert R, Michaelis J (2001) Risk factors for pediatric tumors of the central nervous system: results from a German population-based case-control study. *Med Pediatr Oncol* 36(2):274–282
138. Bunin GR (2004) Nongenetic causes of childhood cancers: evidence from international variation, time trends, and risk factor studies. *Toxicol Appl Pharmacol* 199(2):91–103
139. Ron E, Modan B, Boice JD Jr, Alfandary E, Stovall M, Chetrit A et al (1988) Tumors of the brain and nervous system after radiotherapy in childhood. *N Engl J Med* 319(16): 1033–1039
140. Pacini F, Vorontsova T, Demidchik EP, Molinaro E, Agate L, Romei C et al (1997) Post-chernobyl thyroid carcinoma in Belarus children and adolescents: comparison with naturally occurring thyroid carcinoma in Italy and France. *J Clin Endocrinol Metab* 82(11):3563–3569
141. Whiteman DC, Whiteman CA, Green AC (2001) Childhood sun exposure as a risk factor for melanoma: a systematic review of epidemiologic studies. *Cancer Causes Control* 12(1):69–82
142. International Agency for Research on Cancer (1992) *IARC monographs on the evaluation of carcinogenic risks to humans: solar and ultraviolet radiation*, vol 55. International Agency for Research on Cancer, Lyon, France
143. Vainio H, Bianchini F (2000) Cancer-preventive effects of sunscreens are uncertain. *Scand J Work Environ Health* 26(6):529–531
144. International Agency for Research on Cancer (2004) *IARC monographs on the evaluation of carcinogenic risks to humans: involuntary smoke*, vol 83. International Agency for Research on Cancer, Lyon, France
145. Karube T, Odagiri Y, Takemoto K, Watanabe S (1989) Analyses of transplacentally induced sister chromatid exchanges and micronuclei in mouse fetal liver cells following maternal exposure to cigarette smoke. *Cancer Res* 49(13): 3550–3552
146. Wyrobek AJ, Adler ID (1996) Detection of aneuploidy in human and rodent sperm using FISH and applications of sperm assays of genetic damage in heritable risk evaluation. *Mutat Res* 352(1–2):173–179
147. Pang D, McNally R, Birch JM (2003) Parental smoking and childhood cancer: results from the United Kingdom Childhood Cancer Study. *Br J Cancer* 88(3):373–381
148. Sorahan T, Lancashire RJ (2004) Parental cigarette smoking and childhood risks of hepatoblastoma: OSCC data. *Br J Cancer* 90(5):1016–1018
149. Ji BT, Shu XO, Linet MS, Zheng W, Wacholder S, Gao YT et al (1997) Paternal cigarette smoking and the risk of childhood cancer among offspring of nonsmoking mothers. *J Natl Cancer Inst* 89(3):238–244
150. Sandler DP, Everson RB, Wilcox AJ, Browder JP (1985) Cancer risk in adulthood from early life exposure to parents' smoking. *Am J Public Health* 75(5):487–492
151. Abdollahi M, Ranjbar A, Shadnia S, Nikfar S, Rezaie A, Abdollahi M, Ranjbar A, Shadnia S, Nikfar S, Rezaie A (2004) Pesticides and oxidative stress: a review. *Med Sci Monit* 10(6): RA141–RA147
152. Infante-Rivard C, Weichenthal S (2007) Pesticides and childhood cancer: an update of Zahm and Ward's 1998 review. *J Toxicol Environ Health B Crit Rev* 10(1–2): 81–99

153. Daniels JL, Olshan AF, Teschke K, Hertz-Picciotto I, Savitz DA, Blatt J et al (2001) Residential pesticide exposure and neuroblastoma. *Epidemiology* 12(1):20–27
154. Zahm SH, Ward MH (1998) Pesticides and childhood cancer. *Environ Health Perspect* 106(Suppl 3):893–908
155. Wigle DT, Turner MC, Krewski D, Shannon M, Graef JW (2009) A systematic review and meta-analysis of childhood leukemia and parental occupational pesticide exposure. Lead intoxication from lead-contaminated water used to reconstitute infant formula. *Environ Health Perspect* 117(10):1505–1513
156. Nasterlack M (2006) Do pesticides cause childhood cancer? *Int Arch Occup Environ Health* 79(7):536–544
157. Hill AB (1965) The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine*. 58:295–300
158. Buckley JD, Meadows AT, Kadin ME, Le Beau MM, Siegel S, Robison LL (2000) Pesticide exposures in children with non-Hodgkin lymphoma. *Cancer* 89(11):2315–2321
159. Infante-Rivard C, Sinnott D (1999) Preconceptional paternal exposure to pesticides and increased risk of childhood leukaemia. *Lancet* 354(9192):1819
160. Villanueva CM, Cantor KP, Cordier S, Jaakkola JJ, King WD, Lynch CF et al (2004) Disinfection byproducts and bladder cancer: a pooled analysis. *Epidemiology* 15(3):357–367
161. Boffetta P (2006) Human cancer from environmental pollutants: the epidemiological evidence. *Mutat Res* 608(2):157–162
162. International Agency for Research on Cancer (2004) Some drinking-water disinfectants and contaminants, including arsenic, vol 84. International Agency for Research on Cancer, Lyon, France
163. Infante-Rivard C (2008) Chemical risk factors and childhood leukaemia: a review of recent studies. *Radiat Prot Dosim* 132(2):220–227
164. Raaschou-Nielsen O, Reynolds P (2006) Air pollution and childhood cancer: a review of the epidemiological literature. *Int J Cancer* 118(12):2920–2929
165. Snedeker SM, Diaugustine RP (1996) Hormonal and environmental factors affecting cell proliferation and neoplasia in the mammary gland. *Prog Clin Biol Res* 394:211–253
166. Hildreth NG, Shore RE, Dvoretzky PM (1989) The risk of breast cancer after irradiation of the thymus in infancy. *N Engl J Med* 321(19):1281–1284
167. Herbst AL (1999) Diethylstilbestrol and adenocarcinoma of the vagina. *Am J Obstet Gynecol* 181(6):1576–1578
168. Strohsnitter WC, Noller KL, Hoover RN, Robboy SJ, Palmer JR, Titus-Ernstoff L et al (2001) Cancer risk in men exposed in utero to diethylstilbestrol. *J Natl Cancer Inst* 93(7):545–551
169. Colton T, Greenberg ER, Noller K, Resseguie L, Van Bennekom C, Heeren T et al (1993) Breast cancer in mothers prescribed diethylstilbestrol in pregnancy. Further follow-up. *JAMA* 269(16):2096–2100
170. Olshan AF, Daniels JL (2000) Invited Response: pesticides and childhood cancer. *Am J Epidemiol* 151:647–649

Environmental Toxicology: Oxidative Stress

DEAN P. JONES

Division of Pulmonary, Allergy and Critical Care Medicine, Department of Medicine, Emory University School of Medicine, Atlanta, GA, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Evolution of “Free Radical” and “Radical-free”

Concepts of Oxidative Stress

Radical Mechanisms of Environmental Toxicity

Non-radical Mechanisms of Environmental Toxicity

Future Directions

Acknowledgment

Bibliography

Glossary

Antioxidant A term loosely defined as something that stops oxidation when present at a low amount. The term is used most specifically to mean a free radical-scavenging chemical that accepts or donates an electron to a radical to terminate radical chain reactions. The term is also used for agents that inactivate radical initiation and catalysis, such as metal ion chelators, agents that block radiation-induced oxidation, and agents that counter oxidation of a substance with reduction.

Disulfide An intermediate oxidation state of sulfur that is readily interconverted with the reduced (thiol) state and commonly found in biomolecules. Disulfides are important in protein structure, function, regulation, and translocation. Disulfide formation in proteins and low molecular mass sulfur-containing chemicals (glutathione, cysteine, Co-enzyme A) is a common and sensitive indicator of oxidative stress.

Electrophile (or electrophilic chemical) A chemical with a functional group that is deficient in electron density, such as a quinone or conjugated aldehyde,

and therefore reactive with a nucleophilic chemical containing electron-rich functional groups, such as a thiol. Electrophiles are important in toxicology because they react with nucleophilic centers in DNA, causing mutations and cancer, and with proteins, disrupting enzyme, receptor, transporter, gene expression, and other critical functions.

Lipid peroxidation A radical chain reaction involving polyunsaturated fatty acids (PUFA) and O_2 . The process is initiated by agents that abstract a hydrogen atom from the carbon skeleton to form a carbon-centered radical, propagated by addition of O_2 and abstraction of a hydrogen atom from another carbon skeleton with creation of a lipid hydroperoxide and another radical, and ultimately terminated by reaction of radicals with each other to create non-radical species. The process causes fats and oils to become rancid and is common in decaying organic matter.

Oxidative stress A process or outcome in which an imbalance in prooxidant and antioxidant reactions causes macromolecular damage and/or disruption of biologic redox signaling and control. All classes of macromolecules and any biologic function can be affected. Effects on DNA are important in mutagenesis and cancer, effects on proteins contribute to many acute and chronic toxicities, effects on lipids contribute to dysfunctional energy metabolism and immunity, and effects on carbohydrates affect biological lubricants in joints and cellular identity.

Prooxidant An agent that stimulates aberrant electron transfer in biologic systems and causes oxidative stress. This includes agents that initiate radical reactions, oxidize biologic components, or interfere with normal reductive and antioxidant functions.

Radical (or free radical) An organic molecule with an unpaired electron. Radicals are important in biologic systems because they are reactive and can disrupt biologic functions. Radicals participate in a unique type of chemical reaction sequence termed a chain reaction. A radical chain reaction, such as that illustrated by lipid peroxidation, is initiated by an agent that removes (or adds) an electron from a non-radical chemical to form a radical. The chemical structure is modified, and the radical product becomes a non-radical by accepting or donating an electron to another non-radical species, thereby

propagating the reaction sequence. The process is ultimately terminated by reaction of radicals with each other to create non-radical species.

Reactive oxygen species (ROS), sometimes reactive oxygen intermediates (ROI) One or more oxygen-containing chemicals that are reactive with organic molecules. This term is often used to refer to superoxide anion radical, hydrogen peroxide, lipid hydroperoxide, and related oxygen-centered radicals, but may also include reactive nitrogen species (RNS), such as nitric oxide, peroxyxynitrite, and other oxides of nitrogen (NO_x), which also contain oxygen. The term is most often used because the chemistry is too complex and/or the analytic methods are insufficient to identify specific reactive species.

Redox cycling A process in which a chemical accepts an electron from a biological reductant and transfers that electron to O_2 or other acceptor to create aberrant radical generation. Many quinones and other aromatic chemicals do this by interacting with flavoproteins. The process is considered "cycling" because the chemical functions only as a catalyst. After donating the electron to an acceptor, the original form is regenerated so that it can accept another electron. If electrons are transferred between redox-signaling pathways, this creates a "short circuit," disrupting cell regulation.

Redox signaling A mechanism for cellular communication involving an oxidation–reduction reaction. This process commonly involves generation of a small diffusible redox-active chemical that transfers a signal to a biologic receptor. Redox signaling is highly integrated with kinase signaling and other central cell communication mechanisms, and difficult to discriminate from a more general process of redox sensing, which coordinates cellular functions via widely distributed redox-sensitive cysteine residues in proteins.

Thiol An organic form of sulfur containing a sulfur-hydrogen covalent bond. This is the reduced form of sulfur in the amino acid cysteine, in the antioxidant glutathione, and in most cysteine residues in cellular proteins. In proteins, thiols are also termed "sulfhydryl groups." The thiol functional group supports diverse reactivity and catalytic functions of proteins. Interconversion with disulfides and

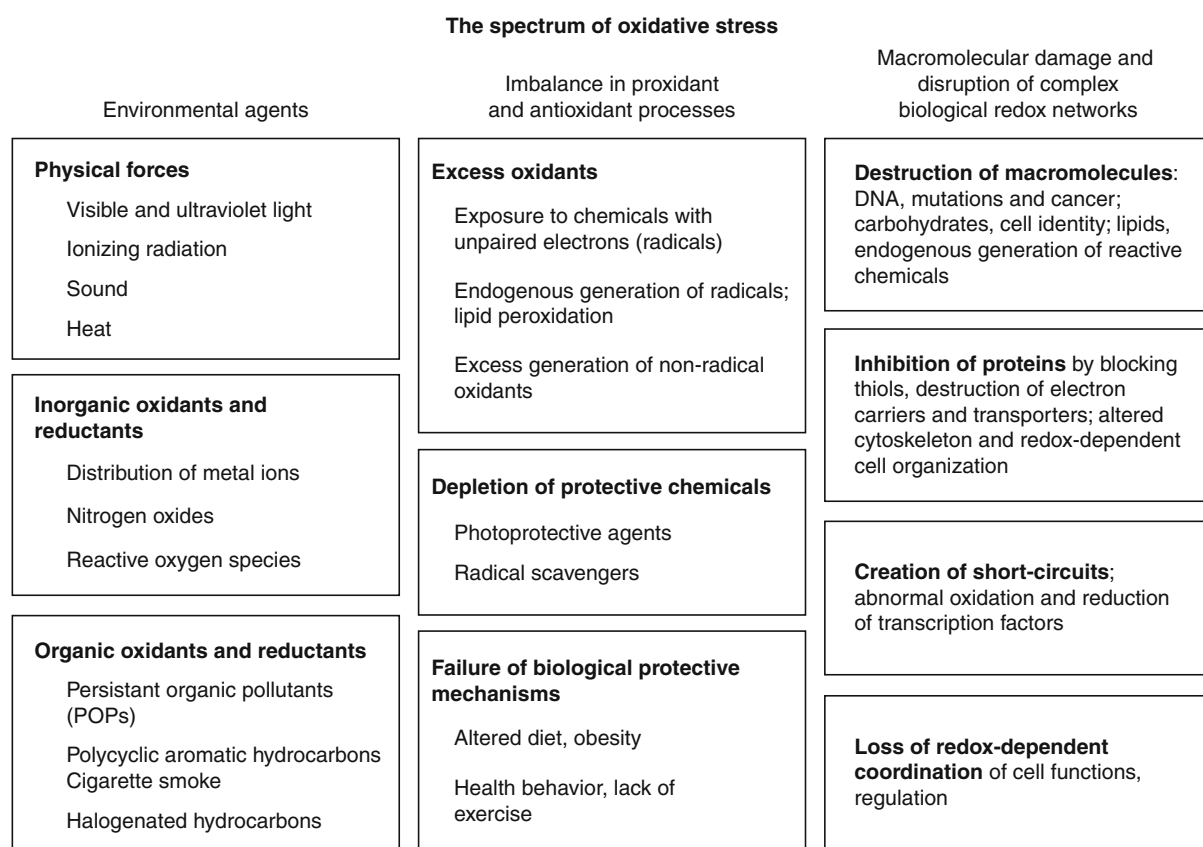
other forms allows thiols to serve as structural transducers as well as chemical signal transducers.

Xenobiotic A chemical that is foreign to an organism. Although often used to refer to man-made chemicals, the term is also used to discriminate chemicals derived exogenously from the environment from those that are generated endogenously as a consequence of intermediary metabolism.

Definition of the Subject

Oxidative stress is a process or outcome in which an imbalance in prooxidant and antioxidant reactions

causes macromolecular damage and/or disruption of biologic redox signaling and control (Fig. 1) [1]. Oxidation–reduction reactions, or “redox” reactions, are reactions involving electron transfer. These reactions are central to energy metabolism and maintenance of metabolic and physical organization of living organisms. Environmental agents that interfere with redox reactions or promote abnormal redox reactions are highly destructive to biologic macromolecules and metabolic organization. Electron transfer reactions in the atmosphere, water, and soil are common and important subjects of environmental sciences. In many



Environmental Toxicology: Oxidative Stress. Figure 1

The spectrum of oxidative stress in environmental toxicology. Oxidative stress refers to adverse effects (*right*) as well as the processes (*center*) that cause adverse effects in biologic systems. Environmental factors that contribute to oxidative stress (*left*) include physical, inorganic, and organic agents. These agents create imbalances in specific prooxidant and antioxidant reactions or reaction pathways (*center*) due to excess oxidants, depletion of protective chemicals, or failure of endogenous protective mechanisms. The imbalances result in the generation of reactive radicals and non-radical oxidants that cause macromolecular damage, inhibit thiol-dependent proteins, short-circuit redox pathways, and disrupt regulatory mechanisms (*right*)

cases, physical processes and inorganic chemicals are relevant causes of oxidative stress to living organisms. Oxidative stress can also occur endogenously within an organism due to organic chemicals derived from the environment. This entry focuses on oxidative stress within living organisms, recognizing that the environmental exposures causing oxidative stress in an organism are highly variable (Fig. 1, left). Many of these are addressed in detail in other sections of this Encyclopedia.

Living organisms depend upon the capability to maintain chemical and physical homeostasis despite variations in environmental exposures. Exposures that cause oxidative stress include chemical oxidants such as ozone and nitrogen oxides of the atmosphere [2, 3], transition metals such as iron [4] and cadmium [5–7], visible and ultraviolet radiation from the sun [8, 9], ionizing radiation from radioactive decay [10, 11], and a large number of pharmaceuticals [12] and chemicals from commercial products and industrial wastes [13, 14]. In some cases there is no real change in the amount of environmental toxicant, only a change in distribution that affects exposure of living organisms [15]. In other cases, there is an accumulation of environmental toxicant due to human activity. Some of these are recognized as persistent organic pollutants (POPs) [16–18] and other times they are persistent products of organic pollutants (P-POPs), some of which are not currently known to accumulate or represent a hazard. For instance, the recent investigation of PFOA, a product of fluorocarbons that is of unknown origin [19, 20], suggests that other unidentified chemicals derived from human activity may exist. Biologic defense mechanisms are diverse and include physical barriers, chemical detoxification mechanisms, repair mechanisms, and avoidance behaviors. Considerable scientific data are available in each of these areas, and efforts are currently being initiated to assemble this knowledge in a systematic description of the “exposome,” i.e., the summation of all environmental exposures from conception throughout the life of an individual [21, 22].

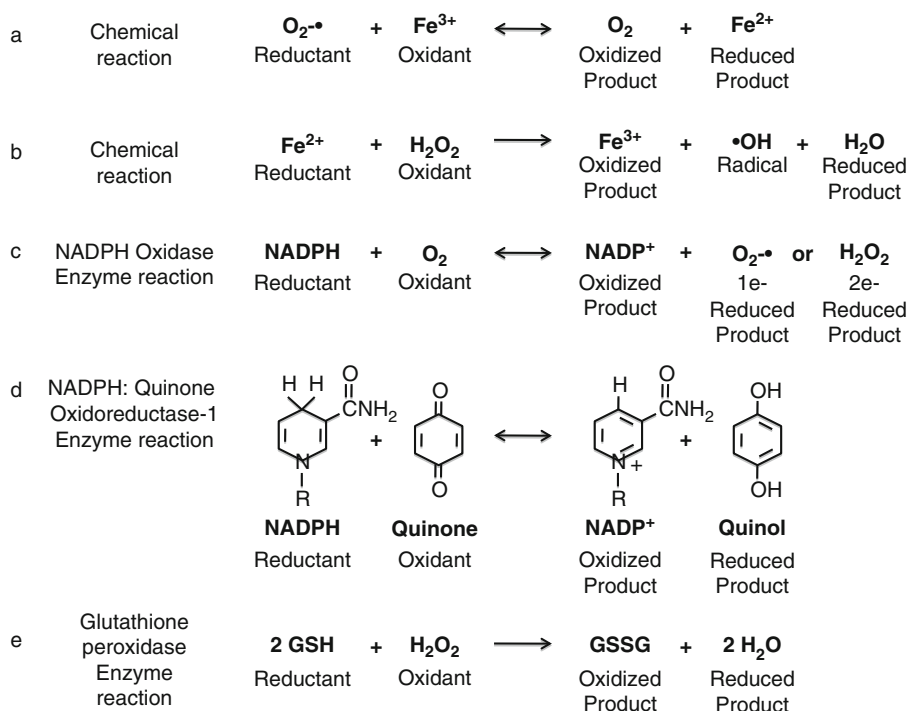
The concepts of oxidative stress are currently undergoing revision from the earlier definition as an imbalance of prooxidants and antioxidants that cause macromolecular damage [23] to one that also includes

oxidative stress as a disruption of vital signaling and control processes within a biologic system [1, 24]. In human health, the accumulating data suggest that the latter is more relevant. However, in environmental toxicology, both macromolecular damage and disruption of redox signaling and control mechanisms are relevant. Importantly, new developments in systems biology [25, 26], including powerful new molecular approaches and analytic tools to study genomics [27] and epigenomics [28], gene expression (transcriptomics) [29], protein modifications (proteomics) [30, 31] and chemical profiles (metabolomics) [32, 33] provide capabilities to improve surveillance, mechanistic understanding, and risk reduction.

Introduction

Two-electron and one-electron transfers in biologic systems. Living organisms obtain the energy necessary to sustain organization, delineation from their external environment, and reproduction, through electron transfer reactions. Loss of electrons is termed “oxidation,” while gain of electrons is termed “reduction” (Fig. 2a–c). Because of the conservation of matter in chemical reactions, the processes are always coupled, so the terms “oxidation–reduction” reactions and “redox” (short for “reduction–oxidation”) reactions are usually more correct. However, the common process of oxidation of chemicals in the presence of atmospheric O₂ (with water as the reduction product) explains the bias in terminology. Excessive reduction can also occur, termed “reductive stress” [34, 35], and is especially relevant to hypoxic and anoxic environmental conditions.

The major elements of life, carbon, hydrogen, oxygen, nitrogen, phosphorus, and sulfur form multiple stable biochemical structures in which pairs of electrons are shared. The major macromolecules (protein, nucleic acids, lipids, carbohydrates) are formed from smaller biochemicals (building blocks: e.g., amino acids, nucleotides, fatty acids, sugars) that include these elements. The biochemical structures are interconverted through chemical reactions that involve conservation of these pairs of electrons. The interconversions are often slow without a catalyst. Biologically, this involves two-electron transfers by NADH or



Environmental Toxicology: Oxidative Stress. Figure 2

Oxidation–reduction reactions. **(a)** One-electron transfer reaction from superoxide anion radical ($\text{O}_2^{\bullet -}$) to Fe^{3+} , as occurs in transfer from superoxide to oxidized cytochrome c. In this reaction, superoxide anion radical is oxidized to molecular oxygen (O_2), while the iron is reduced from Fe^{3+} to Fe^{2+} . Although the reaction can occur in the opposite direction, such as the generation of superoxide by electron transfer in oxy-hemoglobin, energetics in biologic systems usually favor the forward direction as written. **(b)** One-electron transfer from Fe^{2+} to hydrogen peroxide (H_2O_2). This energetically favorable reaction generates three products, Fe^{3+} , hydroxyl radical ($\bullet\text{OH}$), and water. The hydroxyl radical is extremely reactive and destructive to biochemicals. Combination of reaction A and B constitutes the iron-catalyzed Haber–Weiss reaction, also known as the Fenton reaction, which has a central role in radical-mediated macromolecular damage. This process is very destructive when appropriate concentrations of superoxide, H_2O_2 , and iron are present. **(c)** Flavin-containing enzymes, such as NADPH oxidases, catalyze both 1-e[−] and 2-e[−] transfers to O_2 , thereby forming both superoxide and H_2O_2 . Aerobic cells contain NADPH, O_2 , and flavoproteins so that conditions are poised to support Fenton chemistry (A + B) when these enzymes are active. The most potent NADPH oxidase, Nox2, is present in phagocytic cells and activated to kill the invading microorganisms. **(d)** Two-electron transfer, shown for the detoxification enzyme NADPH:quinone reductase-1 (NQO1), is used to support chemical interconversions in intermediary metabolism. The niacinamide group shown for NADPH is identical to that in NADH, and serves to transfer a pair of electrons as a hydride ion (H^-) without radical formation. The reaction shown is an important detoxification reaction by which common environmental quinones are reduced to quinols for phase 2 conjugation and excretion. Quinones can function either as oxidants or electrophiles, and can also generate radicals through redox cycling (see Fig. 3). These adverse reactions are prevented by the NQO1 reaction. **(e)** Thiols are used as reductants to eliminate H_2O_2 . In the glutathione (GSH) peroxidase reaction shown, two molecules of GSH are used to reduce H_2O_2 to water without forming radicals as intermediates. Reduction of the oxidized product, GSSG, by an NADPH-dependent reductase (not shown) completes a detoxification cycle for peroxides. This reaction cycle is complemented by a thioredoxin/peroxiredoxin system to maintain peroxide concentrations very low in cells despite continuous generation. In some cases thiol radicals are formed, but these rapidly react to form disulfides, thereby minimizing toxic effects

NADPH, with the electrons being transferred as a hydride (H^-) ion (Fig. 2c–e). Enzymes serve as catalysts to enhance rates of reactions so that the presence of enzymes largely determines the biochemical pathways for interconversion of the carbon-containing biochemicals. Many of these interconversions involve reaction of “electrophiles,” chemicals with orbitals deficient in electrons, and “nucleophiles,” chemicals with orbitals with excess electron density.

The two-electron (2-e^-) transfer reactions are contrasted by one-electron (1-e^-) transfers, which are predominant reactions of many biologically important metals, such as iron and copper, environmentally important metals, such as cadmium, chromium, and tin, and atmospheric reactions driven by ionizing and non-ionizing radiation. Living organisms are dependent upon 1-e^- chemistry because the central energy capture and transfer reactions of photosynthesis and mitochondrial respiration involve 1-e^- transfers. However, these 1-e^- transfers are physically contained within subcellular structures (chloroplasts and mitochondria), so that both 1-e^- and 2-e^- chemistries can occur simultaneously and be used to provide critical reaction pathways in biological systems. The incompatibility is illustrated by the combination of reactions A and B in Fig. 2, where superoxide donates an electron to Fe^{3+} to form Fe^{2+} , and Fe^{2+} reacts with hydrogen peroxide to form the highly reactive and destructive hydroxyl radical. The 1-e^- systems of mitochondria and chloroplasts are important in environmental toxicology because they represent a loaded gun; these systems have very high electron transfer rates, are sensitive to visible and ultraviolet light, trace metals, oxidants and electrophiles, and cause extensive destruction to other biologic components when triggered by environmental toxicants.

The 2-e^- and 1-e^- pathways are connected through flavoproteins and hydroquinone/quinone structures that exist in interconvertible fully oxidized, 1-e^- -reduced and 2-e^- -reduced forms. Flavoproteins are widely used in biology to support 2-e^- transfer reactions involving NADH and NADPH. In addition, they are critical to connect 2-e^- pathways to 1-e^- pathways, accepting 2e^- from NADH or NADPH and transferring 1e^- to electron acceptors. An illustration is given in Fig. 2c, where different NADPH oxidases support both 1- and 2-e^- reduction of O_2 . Flavoproteins

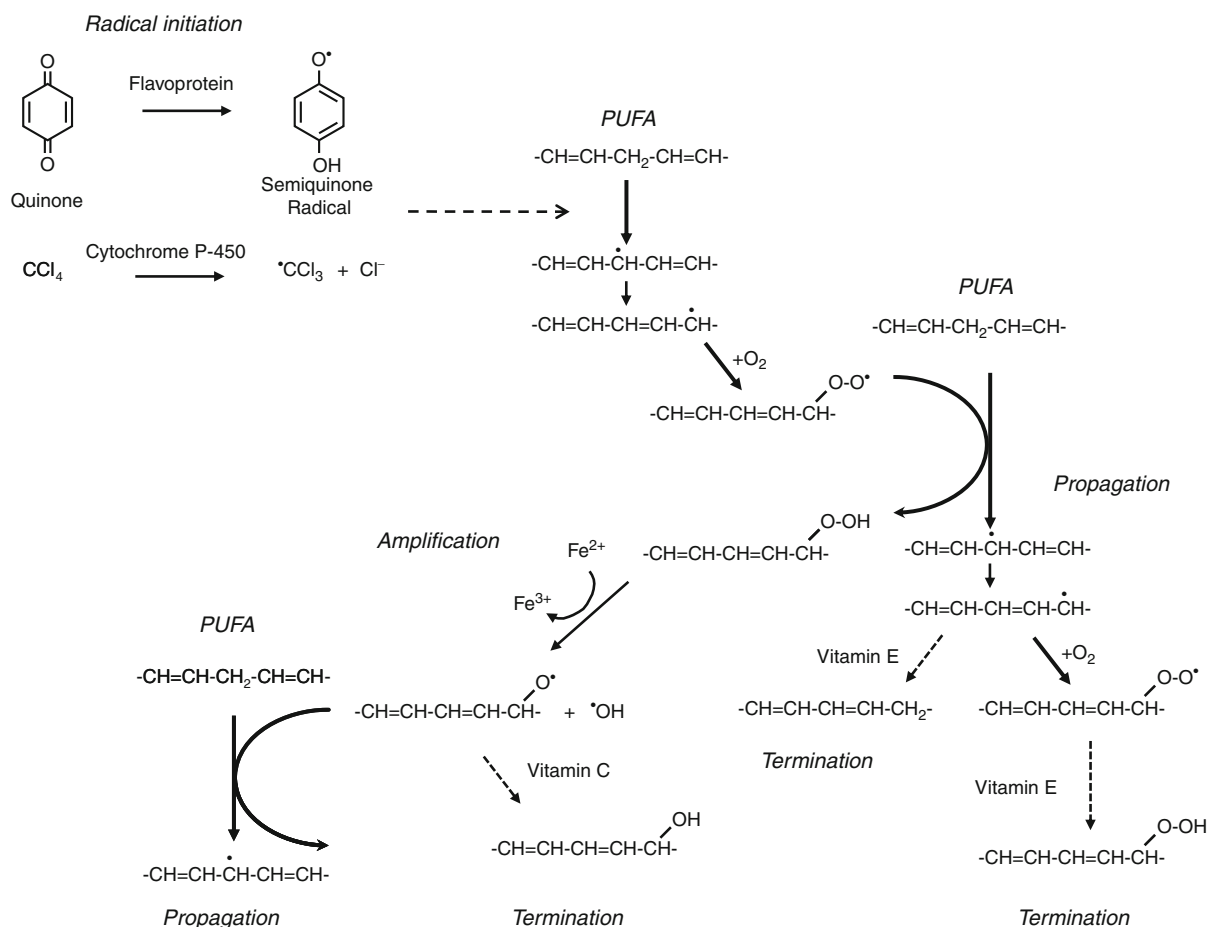
catalyze 1-e^- transfer to the mitochondrial respiratory chain, the major enzyme systems for oxidation of xenobiotics, and to O_2 to form ROS for signaling and defense against invading microorganisms. They also are a major source of toxic oxidant generation by initiating lipid peroxidation (Fig. 3). The planar aromatic structure interacts with polycyclic aromatic structures, facilitating the generation of reactive chemicals with an unpaired electron, termed “radicals,” in a process known as redox cycling [36]. Flavins and flavoproteins are also photoreactive because they absorb visible light, and this activates electrons for transfer to other chemicals.

Antioxidants include chemicals that are relatively stable as radicals and donate or accept an electron to reactive radicals to decrease their destructive chemistry. This is highly beneficial by preventing radical-mediated destruction of macromolecules. Vitamin C (ascorbic acid) and vitamin E (α -tocopherol) are radical-scavenging antioxidants and function to terminate free radical chain reactions (See Fig. 3). Vitamin C is water soluble and synthesized by most animals but not by humans. Vitamin E is fat soluble and one of several lipophilic chemicals that blocks radical reactions in biologic systems. Plants are rich in radical-scavenging antioxidants, and also in carotenoids, light-absorbing chemicals that protect against light-induced oxidative damage.

Evolution of “Free Radical” and “Radical-Free” Concepts of Oxidative Stress

Joseph Priestley, an eighteenth century theologian and chemist credited with the first publication demonstrating production of O_2 [37], intimated that “dephlogisticated air” was toxic even before Lavoisier demonstrated that “dephlogisticated air” was oxygen [38]. Priestley said of breathing dephlogisticated air “. . . I fancied that my breast felt peculiarly light and easy for some time. . .” and “a candle burns out much faster in dephlogisticated air than in common air, so we might, as may be said, live out too fast, and the animal powers be too soon exhausted in the pure kind of air. A moralist, at least, may say, that the air which nature has provided for us is as good as we deserve” [39].

This conclusion of the late eighteenth century presaged a cliché of late twentieth century science, that



Environmental Toxicology: Oxidative Stress. Figure 3

Lipid peroxidation is a common radical chain reaction that occurs in biologic systems. Radical chain reactions include initiation, propagation, and termination events. In biologic systems, initiation (*top left*) often occurs by redox cycling of quinone compounds and reductive activation of halogenated hydrocarbons (e.g., CCl₄). In redox cycling, quinones accept an electron from a reduced flavoprotein, forming a radical. The radical can donate or accept an electron through a hydrogen abstraction from a polyunsaturated fatty acid (PUFA) (*top middle*). Normally this is most energetically favorable at a carbon adjacent to carbons with double bonds. The resulting radical rearranges to form a conjugated double bond system with an adjacent carbon-centered radical. Under aerobic conditions, O₂ rapidly adds to form a peroxyl radical (*center*). The peroxyl radical reacts with another PUFA (*right middle*) to propagate the free radical chain reaction. In this reaction, the peroxyl radical is reduced to a lipid hydroperoxide, which is normally reduced by GSH peroxidases or peroxiredoxin-6 (not shown). When activities of the latter systems are impaired and free iron is present, the process can amplify lipid peroxidation by Fe²⁺-dependent initiation of another radical chain reaction (*bottom left*). In the presence of efficient radical scavengers, such as vitamin C and vitamin E, the chain reactions are terminated. Note that vitamins C and E are most active with different types of radicals and thus complement each other in function. Free radical chain reactions are limited in normal living tissues due to the presence of antioxidants and very high protein concentration. Lipid peroxidation in oils, fats, and decaying matter is very active when antioxidants are depleted.

“oxygen is a double-edged sword.” While science is usually better than cliché, the cliché was highly effective in promoting science, and in this case, an almost religious fervor to introduce “antioxidants” in dietary, cosmetic, and other commercial product development. Failure of large-scale, double-blind interventional trials with antioxidants in humans [40–47] has led to a backlash [48]. Some scientists now have a dismissive attitude concerning the importance of oxidative stress. However, the scientific knowledge of the mechanisms of oxidative stress is very strong and clearly establishes oxidative stress as relevant to environmental toxicology. A brief look at the evolution of concepts facilitates resolution of apparent conflicts between the extensive biomedical literature on oxidative stress and implications for environmental health.

Catalase was one of the first enzymes to be studied [49], and its association with aerobic life was recognized by early biochemists [50]. Hydrogen peroxide (H_2O_2) was recognized as a toxic reactive oxygen species (ROS) associated with O_2 exposure [51]. Early studies also included 1-e^- -transfer reactions of the mitochondrial electron transfer pathway, begun by Keilin in the 1920s [52]. During this time, the environmental agent paraquat (methyl viologen) was introduced as an “electron mediator” to transfer electrons into the mitochondrial pathway [53]. The concepts of oxidative stress, though, are perhaps better traced to the concept of “free radicals,” organic molecules with an unpaired electron, and their damaging effects on biological systems. The chemical terminology for a free radical has been changed to “radical,” but the former is still more commonly used.

The study of radicals in biologic systems was popularized by the discovery of lipid peroxidation (Fig. 3) and antioxidant actions of vitamin E [54] and selenium-dependent glutathione peroxidase [55, 56]. Additional toxicologic studies of carbon tetrachloride (Fig. 3) showed that radical reactions of lipid peroxidation were responsible for hepatotoxicity [57, 58]. Finally, Denham Harmon attracted attention through his hypothesis that free radicals contributed to aging [59].

Lipid peroxidation is a free radical chain reaction contributing to rancidity of oils and fats and invariably present in dead and decaying matter [60]. Radical chain reactions involve an initiation event in which a radical is formed, subsequent propagation events,

and finally termination reactions that eliminate the radicals (Fig. 3). Chemicals, such as quinones, that interact with flavoproteins and redox cycle [36] are a common source of initiation (Fig. 3, top left). Another process involves activation of halogenated hydrocarbons and other environmental chemicals to free radicals by the hemoproteins of the cytochrome P450 (CYP) family (Fig. 3, top left). These radicals abstract a hydrogen atom from polyunsaturated fatty acids (PUFA; Fig. 3, top center), which result in the rearrangement of the newly formed fatty acid radical to a more stable form with conjugated double bonds (Fig. 3, center). O_2 readily adds to this intermediate to form a peroxy radical, which then propagates the chain reaction by the abstraction of a hydrogen atom from another PUFA (Fig. 3, right center). In the presence of trace Fe^{2+} , the process can be amplified by the creation of a new radical from the lipid hydroperoxide (Fig. 3, lower left). Vitamins C and E terminate the process by reducing the reactive intermediates and breaking the chain reaction.

While this constellation of findings and conceptual developments advanced the biochemical study of free radicals, the unanticipated finding of an enzyme that detoxifies superoxide attracted the attention of the broad spectrum of respiratory biologists and invigorated a new field of free radical biology and medicine. Study of the O_2 -dependent reduction of cytochrome *c* by xanthine oxidase led to the discovery that the abundant blood protein erythrocyte protein was an enzyme that converted the 1-e^- reduction product of O_2 , i.e., the superoxide anion radical ($\text{O}_2^{\cdot-}$) to O_2 plus H_2O_2 [61]. The discovery of this enzyme, superoxide dismutase [62], was seminal because it provided evidence that biological systems had evolved an enzyme to eliminate radicals. This provided clear evidence that radicals must be commonly produced and be a threat to biological systems.

Chance et al. [63] provided an authoritative review on hydroperoxides that served as a platform for the integration of concepts of biologic damage by 1-e^- and 2-e^- oxidative processes by Sies [23] in the first monograph entitled “Oxidative Stress.” In this volume, Sies [23] defined oxidative stress as “an imbalance of prooxidants and antioxidants in which the former predominates and causes macromolecular damage.” This clear and simple definition proved useful to stimulate

research on oxidative stress as a mechanism of disease and aging, and antioxidants as means to protect against these deleterious processes.

The contemporary definition of oxidative stress has been refined based upon research findings of the turn of the century [1, 24, 64, 65]. Discovery of a family of enzymes, NADPH oxidases, that generates $O_2^{\cdot-}$ and H_2O_2 as signaling molecules [66, 67] showed that one cannot simply equate ROS with harmful chemistry, i.e., oxidative stress.

At the same time, results of large-scale, double-blind interventional studies with radical-scavenging antioxidants began to appear in the literature [40–47]. These studies showed that supplementation with antioxidants, i.e., shifting the “balance” toward protection, had little or no health benefit in humans. Consequently, the original definition of oxidative stress needs to be qualified in the sense that an “imbalance” is only considered in terms of specific reactions or pathways and not in terms of a global imbalance. In other words, an imbalance in a specific reaction or pathway can cause damage without an overall imbalance in the system, and a change in overall balance of prooxidants and antioxidants can occur without causing damage.

The contemporary view of oxidative stress has also been impacted by the development of new –omic technologies and systems biology. Instead of being limited to overall measures such as cell death [68], lipid peroxidation [69], or protein carbonylation [70], one can now measure effects on expression of specific genes [29], modification of specific lipids [71], and oxidation of specific proteins [72]. This research now shows that oxidative stress as it applies to environmental toxicology involves two processes: one related to earlier concepts of radical-mediated macromolecular damage and the other related to disruption of redox signaling and control mechanisms, which are largely non-radical in nature.

Radical Mechanisms of Environmental Toxicity

The lipid peroxidation mechanisms described above (Fig. 3) are more complex in living organisms because of the high concentrations of protein, the abundance of free radical scavenging antioxidants, and the processes of biologic turnover, which eliminate and replace damaged macromolecules and cells. In addition to redox

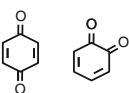
cycling agents and halogenated hydrocarbons, a number of other initiators occur, including transition metals, ionizing and non-ionizing radiation, and sound. Because of the high concentration of protein in cells, the number of propagation reactions tends to be small, emphasizing the importance of initiation events in toxicity.

In pure lipid systems, free radical chain reactions can occur in which an initiation event causes modification of 200–400 molecules of fatty acid prior to being terminated by radicals reacting with each other. However, in biologic systems, such a chain reaction does not occur for several reasons. One is that initiation events are prevented by keeping free metal ion concentrations very low, and blocking other common initiation events. Another is that the concentration of protein is so high that $H\cdot$ is often abstracted from proteins rather than other polyunsaturated lipids, thereby decreasing propagation. Biologic systems also contain high concentrations of radical scavengers, such as vitamin C and vitamin E, which function as chain-terminating antioxidants. Amplification is also effectively prevented by the removal of lipid hydroperoxides by the peroxidases.

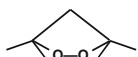
The propagation reactions are also not simple because multiple rearrangement products can occur. In addition to the rearrangement shown in Fig. 3, PUFA radicals can rearrange to eliminate a conjugated aldehyde, such as 4-hydroxynonenal (HNE). HNE is a reactive conjugated aldehyde that can react with proteins or DNA. The reactions generating HNE also generate carbon-centered radicals that propagate chain reactions. A number of other products can be formed, including epoxides (Fig. 4), which can be important as more stable forms with greater specificity in reaction with macromolecules, and isoprostanes [73], rearrangement products that are useful as biomarkers of this chemical reaction sequence. Conjugated aldehydes react with thiols and amines of proteins, creating modified proteins and degradation products that are universally detectable in biologic systems. Thus, lipid peroxidation appears to be an ongoing process. These are measured as protein carbonyls, another useful way to monitor oxidative stress in biologic systems [70].

Lipid peroxides are also formed enzymatically and are metabolized by several peroxidase systems. The most extensively studied enzymes are the glutathione

Ozone	O ₃		
Singlet oxygen	¹ O ₂		
Hydrogen peroxide	H ₂ O ₂		
Hypochlorous acid	HOCl		
Hydroperoxide	ROOH		
Nitrous acid	HNO ₂		
Dinitrogen tetroxide	N ₂ O ₄		
Peroxynitrite	ONOO ⁻		
Alkylperoxynitrite	ROONO		
Disulfide	RSSR'		
Sulfenate	RSO ⁻		
Sulfinate	RSO ₂ ⁻		



Quinones



Endoperoxide



Epoxide

Environmental Toxicology: Oxidative Stress. Figure 4 Non-radical oxidants in biologic systems. Difficulties in quantifying oxidants and their rates of generation, as well as complexities in adaptive and toxic responses to long-term oxidant exposure and a generally conservative approach to terminology, have resulted in widespread use of the term “reactive oxygen species” (ROS). Although convenient, this term does not discriminate between radical and non-radical species and carries a bias that relevant biologic oxidants are closely related to O₂. A large number of non-radical oxidants, some of which are listed in this figure, exist in the environment and in biologic systems. Ozone and singlet oxygen (triplet O₂ is ground state) are common in the atmosphere, while H₂O₂, hypochlorous acid and lipid hydroperoxides are generated enzymatically and non-enzymatically in biologic systems. A number of other reactive carbon oxidation species occur, such as endoperoxides, epoxides and quinones (*right*). Reactive nitrogen oxidation products are also relevant biological oxidants. Nitrous acid and dinitrogen tetroxide are common in the atmosphere and also in biologic systems. Peroxynitrite is a reactive oxidant produced by reaction of two radicals, nitric oxide and superoxide. Nitric oxide (NO[•]) is a protective signaling molecule in the vascular system and is converted to peroxynitrite in the presence of superoxide. Alkylperoxynitrites can similarly form from nitric oxide and lipid peroxyl radicals. Other oxidants can also occur, such as disulfides of cysteine and other low molecular weight thiols. Other sulfur oxidation products, such as sulfenates and sulfinates, can also function as oxidants

(GSH) peroxidases [74], but more recent data indicate that a thioredoxin-dependent peroxidase, peroxiredoxin-6, is more active in the elimination of lipid hydroperoxides [75, 76]. The lipid peroxides can be very important in lipid peroxidation by providing a mechanism for amplification of lipid peroxidation [60]. Trace metal ions, such as Fe²⁺ or Cu¹⁺, can donate an electron to the peroxide to form another radical initiation event.

Because of these processes, radical chain reactions are almost completely prevented in living cells. In contrast, oxidative processes are a component of cellular destruction mechanisms used to kill microorganisms, to establish physical barriers to prevent infection, and to degrade and remove dead and dying cells. Phagocytes have a respiratory burst in which NADPH oxidase-2 generates ROS to kill microorganisms [67]. During apoptosis, a mitochondrial respiratory burst is created by the release of mitochondrial cytochrome *c* and switching the electron transfer chain to one that generates O₂⁻ [77]. Thus, unlike the interpretation that continuous generation of biomarkers of oxidative stress indicates that O₂ is a double-edged sword, the evidence shows that oxidative activities are part of normal biologic processes and that there is reserve capacity to protect against catastrophic collapse. Chain reactions of lipid peroxidation do not provide a major, central toxicologic process, but rather, when lipid peroxidation occurs, it mostly occurs by radical initiation events directly from environmental cause rather than as a consequence of chain reactions.

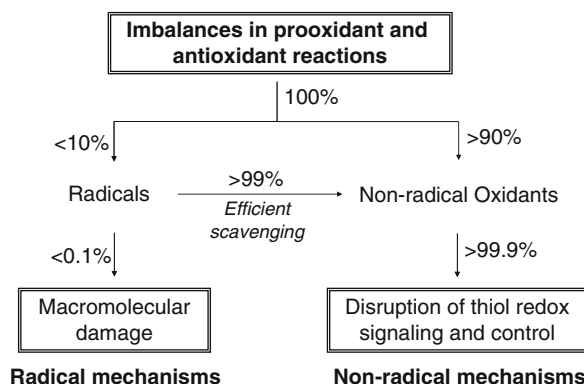
In summary, radical mechanisms are important for environmental toxicology because environmental agents can directly cause initiation of radical reactions that damage all types of macromolecules. Methods are available to evaluate radical-dependent mechanisms of oxidative stress in biological materials, especially through measurement of isoprostanes, protein carbonyls, and other well-documented oxidation products. Low levels of these products are found in normal cells, and healthy organisms are well protected against these reactions. Based upon extensive double-blind interventional studies with free radical scavengers in humans, these reactions do not contribute significantly to long-term health even though many studies show

short-term reductions in biomarkers of oxidative stress. Thus, in the absence of high concentrations of acute environmental exposure, the radical mechanisms appear to be secondary to non-radical oxidative mechanisms of toxicity [78].

Non-radical Mechanisms of Environmental Toxicity

The principles of non-radical mechanisms of oxidative stress are outlined in detail in “Radical-free biology of oxidative stress” [78]. Research has provided extensive knowledge of radical chemistry and antioxidants studied in purified systems and subcellular fractions of biologic systems, often with little regard to biologic relevance [79]. Most critically, biologic systems generate more non-radical oxidants (Fig. 4) than radicals, and the effective radical scavenging processes largely convert radicals to non-radicals or otherwise limit their destructive potential [78]. Important non-radical oxidants (Fig. 4) include H_2O_2 , lipid hydroperoxides, quinones, disulfides, and reactive nitrogen species (RNS), especially peroxynitrite [80]. Peroxynitrite is a major oxidant formed by reaction of superoxide with nitric oxide, a protective signaling molecule controlling microcirculation and immune function [80, 81].

In the xanthine oxidase system, which provided the discovery of superoxide dismutase, the predominant product is H_2O_2 , not the radical superoxide anion [82]. Under optimal conditions for univalent flux to form superoxide, this represents only 30% of the total flux, i.e., under all conditions divalent flux to form H_2O_2 predominated [82]. The univalent flux was considerably less under most conditions, consistent with partitioning of radical and non-radical oxidant production as depicted in Fig. 5. Many redox proteins that generate superoxide show rapid, sequential univalent transfer so that H_2O_2 is a common product of these reactions. Furthermore, dismutation of superoxide anion to H_2O_2 and O_2 occurs at a rapid rate (46) and, as indicated above, reaction of radicals with radical scavengers results in efficient conversion to non-radical oxidants. Consequently, the 2-e^- , non-radical oxidant, H_2O_2 , is produced at a higher rate and appears to represent the major oxidant burden. This rate has



Environmental Toxicology: Oxidative Stress. Figure 5 Non-radical mechanisms of oxidative stress are quantitatively important in chronic and delayed toxicities. Although toxicity due to radical-induced macromolecular damage occurs under some conditions of acute toxicity, accumulating evidence indicates that non-radical mechanisms are important under other conditions. In this figure, the total rate of O_2 reduction to generate reactive oxygen species is expressed as 100%. Two-electron transfers predominate at biologic active sites that reduce O_2 to superoxide and H_2O_2 , so that radicals are always produced at lower rates (e.g., <10% of total) than 2-electron products, such as H_2O_2 (>90% of total). Radical scavengers, including superoxide dismutase, are highly efficient (e.g., >99%) so that the radicals are largely converted to non-radical oxidants. This results in a low level of radical-mediated macromolecular damage and oxidative stress predominantly caused by non-radical oxidants. Non-radical oxidants preferentially target reactive thiols of redox signaling and control pathways, resulting in a disruption of signaling and control. Disruption of signaling and control mechanisms result in long-term dysfunction and reproductive and developmental toxicities. These can be insidious and pose a risk for environmental sustainability

been estimated to be 1–4% of the rate of O_2 consumption in mammalian systems [63], but rates can be higher in some compartments like the peroxisomes [83], or lower, such as in the cytoplasm or nuclei [84].

With non-radical oxidants being quantitatively most important and radical scavenging antioxidants converting some of the radicals to non-radical

oxidants, it is not surprising that radical-scavenging antioxidants do not provide protection against disease [40–47]. Perhaps a better way to view this in terms of environmental toxicity is that radical mechanisms are likely to contribute to acute toxicities but not as likely to contribute to chronic toxicities. In these reactions, macromolecular damage is of major concern, and all classes of macromolecules are potentially vulnerable. On the other hand, non-radical mechanisms can contribute to both; as described below, these are most relevant to chronic toxicity and can have effects through disruption of redox signaling and control, whether or not macromolecular damage occurs.

The major mechanisms of non-radical oxidative stress involve effects on thiols in proteins [78], although mutagenic damage to DNA also occurs [11, 85, 86]. Non-radical oxidants can interfere with reversible oxidation–reduction reactions of thiols by causing abnormal oxidation or irreversible modification. These changes alter physiological function in receptor signaling, transcriptional regulation, cell proliferation, angiogenesis, and apoptosis. Whether or not macromolecular damage occurs, the disruption of these functions alters cell fate and impacts physiologic health. Such changes can be adaptive and protective, in the process known as hormesis [87], but also can cause maladaptation and disease. These processes can occur due to a number of specific environmental agents, can occur in combination, and can be insidious in their long-term effects.

These concepts have been formalized in the redox hypothesis of oxidative stress [78], which provides four postulates to guide investigation into detailed mechanisms, ways to detect oxidative stress, and strategies to intervene to avoid or minimize oxidative stress. These postulates point to a need to incorporate systems biology, modern –omic technologies and bioinformatics into environmental health research, specifically to address environmental toxicities. These four postulates are:

1. All biologic systems contain redox elements (e.g., redox-sensitive cysteine, Cys, residues) that function in cell signaling, macromolecular trafficking, and physiologic regulation.
2. Organization and coordination of the redox activity of these elements occurs through redox circuits

dependent upon common control nodes (e.g., thioredoxin, GSH).

3. The redox-sensitive elements are spatially and kinetically insulated so that “gated” redox circuits can be activated by translocation/aggregation and/or catalytic mechanisms.
4. Oxidative stress is a disruption of the function of these redox circuits caused by specific reaction with the redox-sensitive thiol elements, altered pathways of electron transfer, or interruption of the gating mechanisms controlling the flux through these pathways.

Proteins contain three functional groups that undergo reversible oxidation, the thiol in cysteine (Cys), the thioether in methionine (Met), and the selenol in selenocysteine (Sec). Cys has been studied most extensively, with relevant oxidation states including the thiol ($-SH$), disulfide ($-SS-$), sulfenate ($-SO^-$), sulfinic acid ($-SO_2^-$), and sulfonate ($-SO_3^-$). Thiyl radicals ($-RS\cdot$) generated from thiols in the presence of oxygen-centered radicals [88], as well as other reactive sulfur species [89], can also be considered as toxic species in oxidative stress. Thiyl radicals rapidly react to form disulfides [90], and this may serve as a convergence point between $1-e^-$ and $2-e^-$ events. Relatively unstable sulfenates are converted to disulfides in the presence of thiols, and in some protein structures are stabilized as sulfenamides [91, 92]. Higher oxidation states, sulfinates and sulfonates, are typically not reversible in mammalian systems. A yeast enzyme (sulfiredoxin) reduces sulfinate in peroxiredoxin [93] and could be important in redox signaling mechanisms [94, 95].

Reversible oxidations of Met residues [96, 97] and the less common amino acid, selenocysteine (Sec) [98, 99], can also be important in toxicologic mechanisms. Met oxidation to methionine sulfoxide occurs in oxidative stress and aging [97, 100]. A good example, studied extensively in association with cigarette smoking, is the increased lung rigidity associated with loss of α_1 -antitrypsin inhibitor activity upon Met oxidation [101]. The loss of inhibition of elastase results in damage to lung structures and contributes to obstructive lung disease. Additional research shows that two types of methionine sulfoxide reductases (MSR) are important in protecting against distinct S- and

R-sulfoxides in methionine residues [96, 102, 103]. These MSR activities are dependent upon thioredoxins [103] and are associated with longevity [104–107]. This less frequently studied type of oxidation involves both Met and Sec, and probably deserves much greater attention in environmental oxidative stress due to the variable environmental availability of Met and Sec. The selenol of Sec undergoes reversible oxidation–reduction during catalytic functions of thioredoxin reductases and Se-dependent glutathione peroxidases [98, 99], enzymes present at key positions in both thioredoxin and GSH pathways and critically important to protect against non-radical oxidative stress.

Key points of oxidative stress according to the redox hypothesis include recognition of mechanisms of oxidative stress other than an imbalance of prooxidants and antioxidants. One of these involves altered function of redox circuits caused by specific reaction with redox-sensitive thiol elements (Fig. 6). This differs from the earlier concepts of macromolecular damage in that it could affect adaptability or tolerance of an organism without overt toxicity. For instance, redox-sensing thiols could govern tolerance to cold or prolonged starvation. Modification of these thiols would have no effect without the environmental challenge but would have catastrophic outcome with cold exposure or starvation. Environmental exposures could also create new pathways for electron transfer, effectively creating short circuits. Such short circuits may be accommodated by the stability of a network structure but create an overall vulnerability to other exposures.

Finally, a redox system design using low-flux systems to control high-flux systems is economical in terms of energy utilization [64]. Such a structure can be highly sensitive to low level exposures because it allows toxic agents to interrupt the gating mechanisms that control flux through these pathways. Because of this, accumulation of low levels of certain agents may induce a loss of function.

Future Directions

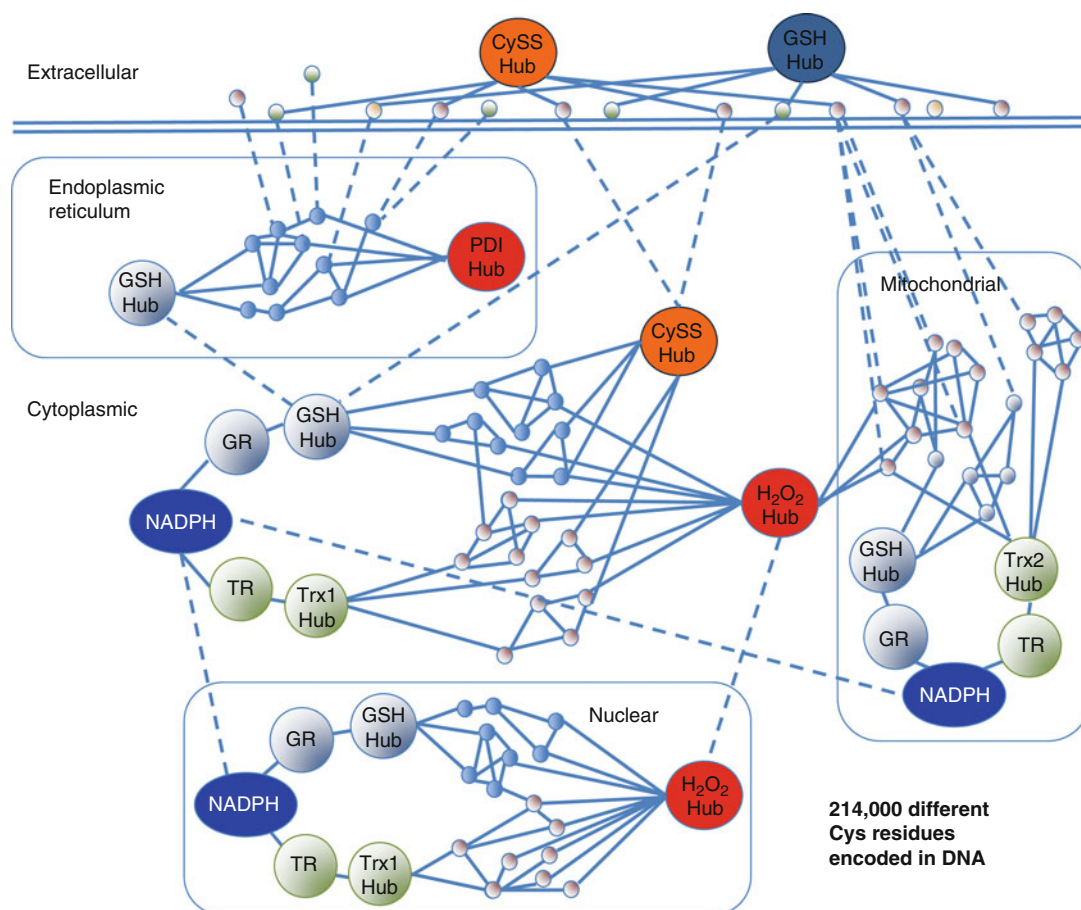
Oxidative stress and environmental sustainability. The view of oxidative stress as a disruption of redox signaling and control, as opposed to an imbalance of prooxidants and antioxidants, has different

implications for sustainability in the environment. If oxidative stress were simply an imbalance, then oxidative stress caused by environmental agents could be eliminated by removing the source or directly compensated by increased antioxidants. Cleanup efforts generally approach the problem as though eliminating the source were sufficient. However, this simplistic approach, like the antioxidant supplementation trials, does not effectively address the problem with regard to environmental sustainability.

A critical issue is that complex systems respond to challenges with adaptation to maintain function. Because of this, subtoxic exposures can accumulate without overt signs of adverse effects. In the long term, this is not sustainable. Consequently, studies to evaluate impact can provide reassuring evidence that specific environmental challenges are not a problem, when the data may really mean that the biologic systems are effectively engineered to conceal hazards. While this may seem overly paranoid, millions of new chemicals have been synthesized in research efforts, and tens of thousands are in commercial use. Consequently, to consider sustainability in terms of oxidative stress, one must evaluate sustainability in terms of the impacts of cumulative exposures and acknowledge the reality that one cannot correct oxidative stress simply by shifting the balance with antioxidants.

Unfortunately, the scientific method is not well designed to deal with complex systems. The range of responses to multiple exposures within complex systems over long periods of exposure cannot be evaluated or predicted accurately with contemporary methods. Sustainability in the environment necessitates new ways to evaluate exposures and biologic responses.

From the standpoint of oxidative stress, a redox proteomics approach provides means for systems analysis. With a currently available liquid chromatography-mass spectrometry approach, one can measure the percent oxidation of specific cysteine residues in hundreds of proteins [108]. These Cys residues represent only a fraction of the total Cys encoded by the genome, which is about 200,000 in the 18,000 proteins in humans, but none-the-less provide an indication of the steady-state oxidation of the individual. The Cys residues change in oxidation state in response to exposures so that one could, in principle, establish sentinel species to examine environmental oxidative stress.



Environmental Toxicology: Oxidative Stress. Figure 6

Targets of oxidative stress include redox-sensitive thiols functioning in subcellular compartments. The human genome encodes 214,000 different cysteine residues. The resulting cysteine proteome is distributed among cellular compartments, with extracellular protein thiols (indicated as nodes) interacting with free cysteine and cystine, designated as a CySS redox-control hub, in the plasma and interstitial space. Solid connectors (edges) indicate direct interactions in a network structure. On the alveolar surface of the lung and lining of the gastrointestinal tract, GSH and GSSG appear to serve as the predominant redox-control hub. The extracellular thiol/disulfide systems are connected to intracellular systems (designated by broken lines) principally through transport systems or other indirect mechanisms. Within the cytoplasm, a CySS hub and a H_2O_2 hub function in the oxidation of cytoplasmic proteins. These oxidative reactions are ongoing and are countered by two major NADPH-dependent thiol reduction systems, represented by thioredoxin-1 (Trx1) and GSH/GSSG hubs. The relative rates for different proteins vary, with the majority of Cys residues being largely reduced under steady-state conditions. Trx1 is translocated into nuclei during oxidative stress; in the nuclei, Trx1 is maintained at a more reduced state than in the cytoplasm. Both nuclear Trx1 and nuclear GSSG hubs function are at least partially independent of respective cytoplasmic hubs. These hubs are important for redox control of transcription and are, therefore, critical nuclear targets of oxidative stress. Mitochondria contain Trx2 and GSH/GSSG hubs that control redox states of mitochondrial proteins. Lipophilic cationic oxidants can be especially toxic because they accumulate in mitochondria and disrupt bioenergetic control. In the secretory pathway, protein disulfide isomerase (PDI) is oxidized by an endoplasmic reticulum oxidase system (EROS) and functions as a redox hub controlling Cys oxidation in proteins destined for the plasma membrane or extracellular compartments. The range of organelle function in specialized tissues results in organ system-specific oxidative stress.

Two types of data can be obtained, one giving maps of electron transfer circuits and the other giving functional pathway maps linked by Cys oxidation [108]. Such a biomonitoring approach could be used to establish cumulative databases, providing a strategy to quantify deterioration of environmental conditions in terms of disruption of protein redox systems in the sentinel species. Of course, humans could be one of the sentinel species, and evaluation of protein oxidation in human samples could be useful to document oxidative stress. Similar information could be obtained from gene expression profiling; complete genomic information is available now for several species, so that multiple options are available to test the feasibility of such approaches.

Unfortunately, changes in redox proteomics or gene expression only reveal whether deterioration has occurred, and do not provide insight into the underlying causes. Exposure information will be needed to link environmental deterioration to specific causes. Current environmental surveillance methods are poorly equipped to address this problem because analyses are highly targeted to a relatively small number of hazardous agents. Newer methods are available to profile chemicals in biologic extracts, which show that thousands of chemicals can be detected in a single analysis [33]. Less than half of the chemicals detected in human plasma have been identified, meaning that humans are exposed to a large number of unidentified chemicals. Some of these may be environmental degradation products of known hazardous and nonhazardous chemicals. Whether these chemicals are helpful or harmful is not known, but multivariate analysis and modern bioinformatic methods can support studies linking alteration in metabolic characteristics with abundance of specific chemicals in biologic samples. Clearly, investment in new methods as well as cumulative environmental health data registries will be needed to address critical questions of environmental sustainability with regard to oxidative stress.

During the past decade, methods have been developed to allow study of oxidative stress within specific subcellular compartments [84, 109]. Mammalian cells contain about 214,000 Cys residues encoded in the genome, and redox biochemical approaches are beginning to provide an understanding of the organization of the redox network structure [108, 110, 111]. This

research has revealed that oxidative stress is not uniform within cells but rather selectively affects specific pathways in compartments. The extracellular compartments of plasma and interstitial space are generally more oxidized than cellular compartments (Fig. 6) [84], have fewer antioxidant systems, and are vulnerable to oxidants. The lung alveoli, oral mucosa, and intestinal lumen are supplied with the thiol antioxidant GSH, and have protective enzymes associated with the mucous-lining layer. Thus, the maintenance of these surface barriers is critical for the protection of an organism because the lung lining is first exposed to airborne toxicants, while the intestine is first exposed to ingested toxicants.

Within cells, different redox systems function in the organelles. The endoplasmic reticulum (ER) and secretory pathway use protein disulfide isomerase (PDI) and an oxidase system (EROS) to oxidize proteins during processing for secretion ([112]; Fig. 6). Disruption of this oxidative pathway activates cell death through an ER stress mechanism [112, 113]. Nuclei and mitochondria are more reduced [84], and each has specific redox-sensitive proteins. Mitochondria have a unique thioredoxin-2, Trx2, and glutaredoxin-2, Grx2, while thioredoxin-1 and glutaredoxin-1 are found in the cytoplasm. Trx1 is translocated into nuclei during oxidative stress [114], and nuclei also contain Grx2, at least when over-expressed in cells [110]. Accumulating evidence indicates that most agents that cause oxidative stress do so by disrupting specific redox control pathways associated with specific organelles (Fig. 6).

Finally, there is a possibility that artificial intelligence may provide a more cost-effective means to monitor environmental sustainability. Irregularity of biologic data often reflects adaptability and health of the system. In essence, the accommodation of oxidative challenges can be expected to decrease the flexibility of the system. Thus, if appropriate methods were developed, it may be possible to evaluate health (or sustainability of a system) in terms of irregularity of metabolic or protein data [115]. Such an approach could potentially be accomplished using monofractality or multifractality, such as expressed by Hurst or Holder exponents [116]. While these are not traditional approaches used by experimental or environmental biologists, they could be very useful to approach the rather difficult problem of environmental

sustainability in the face of such a large number of newly introduced and accumulating abundance of environmental agents.

In summary, oxidative stress is an important concern in environmental toxicology, both as a cause of acute toxicity and as a contributor to chronic toxicity. In terms of sustainability, agents causing acute toxicity due to oxidative stress are more readily identified and managed by containment and cleanup, in some cases complemented by use of metal chelators and free radical scavenging antioxidants. Agents that cause chronic toxicity or delayed reproductive or developmental toxicity are of greater concern because of the difficulty of identification and association with the toxicity. At present, there is no way to evaluate whether this is a minor or major problem, but halogenated hydrocarbons are commonly used. Some of these could accumulate in the biosphere, and most of these are likely to contribute to oxidative stress. Safety testing mostly is performed only on commercial materials and not on products formed by the modification of these materials after release into the environment. Mass spectrometry analyses of human plasma reveal thousands of unidentified ions with unknown biologic effects. Currently there is no environmental exposure census to reveal whether or not such chemicals are accumulating due to human activity. Thus, there is considerable opportunity and need to enhance understanding of the environmental load of chemicals that could accumulate and cause adverse effects due to oxidative stress [117–122].

Acknowledgment

Research support provided by grants from the National Institutes of Environmental Health Sciences ES009047 and ES011195.

Bibliography

- Sies H, Jones DP (2007) Oxidative stress. In: Fink G (ed) *Encyclopedia of stress*. Elsevier, San Diego, pp 45–48
- Ruidavets JB, Cournot M, Cassadou S, Giroux M, Meybeck M, Ferrieres J (2005) Ozone air pollution is associated with acute myocardial infarction. *Circulation* 111:563–569
- Seinfeld JH, Pandis SN (2006) *Atmospheric chemistry and physics – from air pollution to climate change*, 2nd edn. Wiley, New York, 1225 pp
- Becana M, Moran JF, Iturbe-Ormaetxe I (1998) Iron-dependent oxygen free radical generation in plants subjected to environmental stress: toxicity and antioxidant protection. *Plant Soil* 201:137–147
- Das P, Samantaray S, Rout GR (1997) Studies on cadmium toxicity in plants: a review. *Environ Pollut* 98:29–36
- Klaassen CD, Liu J, Choudhuri S (1999) Metallothionein: an intracellular protein to protect against cadmium toxicity. *Annu Rev Pharmacol Toxicol* 39:267–294
- Ragunathan N, Dairou J, Sanfins E, Busi F, Noll C, Janel N, Dupret JM, Rodrigues-Lima F (2010) Cadmium alters the biotransformation of carcinogenic aromatic amines by arylamine n-acetyltransferase xenobiotic-metabolizing enzymes: molecular, cellular, and in vivo studies. *Environ Health Perspect* 118:1685–1691
- Hideg E, Vass I (1996) Uv-b induced free radical production in plant leaves and isolated thylakoid membranes. *Plant Sci* 115:251–260
- Jurkiewicz BA, Buettner GR (1996) Epr detection of free radicals in uv-irradiated skin: mouse versus human. *Photochem Photobiol* 64:918–922
- Bernhard WA (2009) Radical reaction pathways initiated by direct energy deposition in DNA by ionizing radiation. In: Greenberg MM (ed) *Radical and radical ion reactivity in nucleic acid chemistry*. Wiley, Hoboken, pp 41–68
- Jackson SP, Bartek J (2009) The DNA-damage response in human biology and disease. *Nature* 461:1071–1078
- Tafazolli S, Spehar DD, O'Brien PJ (2005) Oxidative stress mediated idiosyncratic drug toxicity. *Drug Metab Rev* 37:311–325
- Araujo JA, Barajas B, Kleinman M, Wang X, Bennett BJ, Gong KW, Navab M, Harkema J, Sioutas C, Lulis AJ, Nel AE (2008) Ambient particulate pollutants in the ultrafine range promote early atherosclerosis and systemic oxidative stress. *Circ Res* 102:589–596
- Siddique HR, Gupta SC, Mitra K, Bajpai VK, Mathur N, Murthy RC, Saxena DK, Chowdhuri DK (2008) Adverse effect of tannery waste leachates in transgenic drosophila melanogaster: role of Ros in modulation of Hsp70, oxidative stress and apoptosis. *J Appl Toxicol* 28:734–748
- Matschullat J, Birmann K, Borba RP, Ciminelli V, Deschamps EM, Figueiredo BR, Gabrio T, Haßler S, Hilscher A, Junghänel I, de Oliveira N, Raßbach K, Schmidt H, Schwenk M, de Oliveira Vilhena MJ, Weidner U, Abmjbrz BP, Richard HL (2007) Long-term environmental impact of arsenic-dispersion in Minas Gerais, Brazil. In: Bundschuh J, Zevenhoven R, Loeppert RH (eds) *Trace metals and other contaminants in the environment*. Elsevier, Great Lakes, pp 365–382
- Jones KC, de Voogt P (1999) Persistent organic pollutants (pops): state of the science. *Environ Pollut* 100:209–221
- Ruzzin J, Petersen R, Meugnier E, Madsen L, Lock EJ, Lillefosse H, Ma T, Pesenti S, Sonne SB, Marstrand TT, Malde MK, Du ZY, Chavey C, Fajas L, Lundebye AK, Brand CL, Vidal H, Kristiansen K, Froyland L (2010) Persistent organic pollutant exposure leads to insulin resistance syndrome. *Environ Health Perspect* 118:465–471
- Wania F, Mackay D (1996) Tracking the distribution of persistent organic pollutants. *Environ Sci Technol* 30:A390–A396

19. Apelberg BJ, Witter FR, Herbstman JB, Calafat AM, Halden RU, Needham LL, Goldman LR (2007) Cord serum concentrations of perfluorooctane sulfonate (PFOS) and perfluorooctanoate (PFOA) in relation to weight and size at birth. *Environ Health Perspect* 115:1670–1676
20. Wallington TJ, Hurley MD, Xia J, Wuebbles DJ, Sillman S, Ito A, Penner JE, Ellis DA, Martin J, Mabury SA, Nielsen OJ, Sulbaek Andersen MP (2006) Formation of C7F15COOH (PFOA) and other perfluorocarboxylic acids during the atmospheric oxidation of 8:2 fluorotelomer alcohol. *Environ Sci Technol* 40:924–930
21. Rappaport SM (2010) Implications of the exposome for exposure science. *J Expo Sci Environ Epidemiol* 21:5–9
22. Wild CP (2005) Complementing the genome with an “Exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomark Prev* 14:1847–1850
23. Sies H (1985) Oxidative stress: introductory remarks. In: Sies H (ed) *Oxidative stress*. Academic, London, pp 1–8
24. Jones DP (2006) Redefining oxidative stress. *Antioxid Redox Signal* 8:1865–1879
25. Voit E, Neves AR, Santos H (2006) The intricate side of systems biology. *Proc Natl Acad Sci USA* 103:9452–9457
26. Voit EO, Qi Z, Miller GW (2008) Steps of modeling complex biological systems. *Pharmacopsychiatry* 41(Suppl 1):S78–S84
27. Schuster SC (2008) Next-generation sequencing transforms today's biology. *Nat Meth* 5:16–18
28. Callinan PA, Feinberg AP (2006) The emerging science of epigenomics. *Hum Mol Genet* 15(Spec No 1):R95–R101
29. Bammler T, Beyer RP, Bhattacharya S, Boorman GA, Boyles A, Bradford BU, Bumgarner RE, Bushel PR, Chaturvedi K, Choi D, Cunningham ML, Deng S, Dressman HK, Fannin RD, Farin FM, Freedman JH, Fry RC, Harper A, Humble MC, Hurban P, Kavanagh TJ, Kaufmann WK, Kerr KF, Jing L, Lapidus JA, Lasarev MR, Li J, Li YJ, Lobenhofer EK, Lu X, Malek RL, Milton S, Nagalla SR, O'Malley JP, Palmer VS, Pattee P, Paules RS, Perou CM, Phillips K, Qin LX, Qiu Y, Quigley SD, Rodland M, Rusyn I, Samson LD, Schwartz DA, Shi Y, Shin JL, Sieber SO, Slifer S, Speer MC, Spencer PS, Sproles DI, Swenberg JA, Suk WA, Sullivan RC, Tian R, Tennant RW, Todd SA, Tucker CJ, Van Houten B, Weis BK, Xuan S, Zarbl H (2005) Standardizing global gene expression analysis between laboratories and across platforms. *Nat Meth* 2:351–356
30. Li H, Grigoryan H, Funk WE, Lu SS, Rose S, Williams ER, Rappaport SM (2010) Profiling Cys34 adducts of human serum albumin by fixed-step selected reaction monitoring. *Mol Cell Proteomics* (Ahead of print)
31. Funk WE, Li H, Iavarone AT, Williams ER, Riby J, Rappaport SM (2010) Enrichment of cysteinyl adducts of human serum albumin. *Anal Biochem* 400:61–68
32. Evans AM, DeHaven CD, Barrett T, Mitchell M, Milgram E (2009) Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Anal Chem* 81:6656–6667
33. Johnson JM, Yu T, Strobel FH, Jones DP (2010) A practical approach to detect unique metabolic patterns for personalized medicine. *Analyst* 135:2864–2870
34. Namakkal Soorappan R, Varadharaj S, Khanderao GD, Davidson CJ, Kannan S, Firpo MA, Zweier J, Benjamin I (2010) Sustained activation of nrf2/are signaling promotes reductive stress in the human mutant protein aggregation cardiomyopathy in mice. *Antioxid Redox Signal* (Ahead of print)
35. Zhang X, Min X, Li C, Benjamin IJ, Qian B, Ding Z, Gao X, Yao Y, Ma Y, Cheng Y, Liu L (2010) Involvement of reductive stress in the cardiomyopathy in transgenic mice with cardiac-specific overexpression of heat shock protein 27. *Hypertension* 55:1412–1417
36. Kappus H, Sies H (1981) Toxic drug effects associated with oxygen metabolism: redox cycling and lipid peroxidation. *Experientia* 37:1233–1241
37. Priestley J (1775) Letter to sir john pringle. *Philos Trans* 65:384–390
38. Bitterman N, Bitterman H (2006) Oxygen toxicity. In: Mathiew D (ed) *Handbook on hyperbaric medicine*. Dordrecht, Springer
39. Silverman WA (2004) A cautionary tale about supplemental oxygen: the albatross of neonatal medicine. *Pediatrics* 113:394–396
40. The Alpha-Tocopherol, Beta Carotene Cancer Prevention Study Group (1994) The effect of vitamin E and beta carotene on the incidence of lung cancer and other cancers in male smokers. *N Engl J Med* 330:1029–1035
41. Gruppo Italiano Per Lo Studio Della Sopravvivenza Nell'infarto Miocardico (1999) Dietary supplementation with n-3 polyunsaturated fatty acids and vitamin e after myocardial infarction: results of the gissi-prevenzione trial. *Lancet* 354:447–455
42. Age-Related Eye Disease Study Research Group (2001) A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E, beta carotene, and zinc for age-related macular degeneration and vision loss: AREDS Report no. 8. *Arch Ophthalmol* 119:1417–1436
43. de Gaetano G (2001) Low-dose aspirin and vitamin e in people at cardiovascular risk: a randomised trial in general practice. Collaborative group of the primary prevention project. *Lancet* 357:89–95
44. Hennekens CH, Buring JE, Manson JE, Stampfer M, Rosner B, Cook NR, Belanger C, LaMotte F, Gaziano JM, Ridker PM, Willett W, Peto R (1996) Lack of effect of long-term supplementation with beta carotene on the incidence of malignant neoplasms and cardiovascular disease. *N Engl J Med* 334:1145–1149
45. Lonn E, Bosch J, Yusuf S, Sheridan P, Pogue J, Arnold JM, Ross C, Arnold A, Sleight P, Probstfield J, Dagenais GR (2005) Effects of long-term vitamin E supplementation on cardiovascular events and cancer: a randomized controlled trial. *J Am Med Assoc* 293:1338–1347
46. Yusuf S, Dagenais G, Pogue J, Bosch J, Sleight P (2000) Vitamin e supplementation and cardiovascular events in high-risk patients. The heart outcomes prevention evaluation study investigators. *N Engl J Med* 342:154–160

47. Zureik M, Galan P, Bertrais S, Mennen L, Czernichow S, Blacher J, Ducimetiere P, Hercberg S (2004) Effects of long-term daily low-dose supplementation with antioxidant vitamins and minerals on structure and function of large arteries. *Arterioscler Thromb Vasc Biol* 24:1485–1491
48. Williams KJ, Fisher EA (2005) Oxidation, lipoproteins, and atherosclerosis: which is wrong, the antioxidants or the theory? *Curr Opin Clin Nutr Metab Care* 8:139–146
49. Loew O (1901) Catalase, a new enzyme of general occurrence. U.S. Department of Agriculture Report no. 68, pp 47–55
50. Chelikani P, Fita I, Loewen PC (2004) Diversity of structures and properties among catalases. *Cell Mol Life Sci* 61:192–208
51. Maehly AC, Chance B (1954) The assay of catalases and peroxidases. *Meth Biochem Anal* 1:357–424
52. Keilin D (1925) On cytochrome, a respiratory pigment, common to animals, yeast, and higher plants. *Proc R Soc Lond B – Containing Pap Biol Character* 98:312–339
53. Dutton PL, Wilson DF (1974) Redox potentiometry in mitochondrial and photosynthetic bioenergetics. *Biochim Biophys Acta* 346:165–212
54. Wolf G (2005) The discovery of the antioxidant function of vitamin E: the contribution of Henry A. Mattill. *J Nutr* 135:363–366
55. Mills GC (1957) Hemoglobin catabolism. I. Glutathione peroxidase, an erythrocyte enzyme which protects hemoglobin from oxidative breakdown. *J Biol Chem* 229:189–197
56. Rotruck JT, Pope AL, Ganther HE, Swanson AB, Hafeman DG, Hoekstra WG (1973) Selenium: biochemical role as a component of glutathione peroxidase. *Science* 179:588–590
57. Recknagel RO, Ghoshal AK (1966) Lipoperoxidation of rat liver microsomal lipids induced by carbon tetrachloride. *Nature* 210:1162–1163
58. Recknagel RO, Ghoshal AK (1966) Lipoperoxidation as a vector in carbon tetrachloride hepatotoxicity. *Lab Invest* 15:132–148
59. Harman D (1956) Aging: a theory based on free radical and radiation chemistry. *J Gerontol* 11:298–300
60. Tribble DL, Aw TY, Jones DP (1987) The pathophysiological significance of lipid peroxidation in oxidative cell injury. *Hepatology* 7:377–386
61. Fridovich I (1997) Superoxide anion radical ($O_2^{\bullet-}$), superoxide dismutases, and related matters. *J Biol Chem* 272:18515–18517
62. McCord JM, Fridovich I (1969) Superoxide dismutase. An enzymic function for erythrocuprein (hemocuprein). *J Biol Chem* 244:6049–6055
63. Chance B, Sies H, Boveris A (1979) Hydroperoxide metabolism in mammalian organs. *Physiol Rev* 59:527–605
64. Jones DP (2006) Disruption of mitochondrial redox circuitry in oxidative stress. *Chem Biol Interact* 163:38–53
65. Jones DP (2006) Extracellular redox state: refining the definition of oxidative stress in aging. *Rejuvenation Res* 9:169–181
66. Griending KK (2006) NADPH oxidases: new regulators of old functions. *Antioxid Redox Signal* 8:1443–1445
67. Lambeth JD (2004) Nox enzymes and the biology of reactive oxygen. *Nat Rev Immunol* 4:181–189
68. Orrenius S (2004) Mitochondrial regulation of apoptotic cell death. *Toxicol Lett* 149:19–23
69. Morrow JD, Roberts LJ 2nd (1999) Mass spectrometric quantification of F2-isoprostanes in biological fluids and tissues as measure of oxidant stress. *Meth Enzymol* 300:3–12
70. Levine RL, Garland D, Oliver CN, Amici A, Climent I, Lenz AG, Ahn BW, Shaltiel S, Stadtman ER (1990) Determination of carbonyl content in oxidatively modified proteins. *Meth Enzymol* 186:464–478
71. Kagan VE, Borisenko GG, Tyurina YY, Tyurin VA, Jiang J, Potapovich AI, Kini V, Amoscato AA, Fujii Y (2004) Oxidative lipidomics of apoptosis: redox catalytic interactions of cytochrome c with cardiolipin and phosphatidylserine. *Free Radic Biol Med* 37:1963–1985
72. Go YM, Park H, Koval M, Orr M, Reed M, Liang Y, Smith D, Pohl J, Jones DP (2010) A key role for mitochondria in endothelial signaling by plasma cysteine/cystine redox potential. *Free Radic Biol Med* 48:275–283
73. Morrow JD, Roberts LJ 2nd (2002) Mass spectrometric quantification of F2-isoprostanes as indicators of oxidant stress. *Meth Mol Biol* 186:57–66
74. Brigelius-Flohe R (1999) Tissue-specific functions of individual glutathione peroxidases. *Free Radic Biol Med* 27:951–965
75. Fisher AB, Dodia C, Manevich Y, Chen JW, Feinstein SI (1999) Phospholipid hydroperoxides are substrates for non-selenium glutathione peroxidase. *J Biol Chem* 274:21326–21334
76. Wang Y, Phelan SA, Manevich Y, Feinstein SI, Fisher AB (2006) Transgenic mice overexpressing peroxiredoxin 6 show increased resistance to lung injury in hyperoxia. *Am J Respir Cell Mol Biol* 34:481–486
77. Cai J, Jones DP (1998) Superoxide in apoptosis. Mitochondrial generation triggered by cytochrome C loss. *J Biol Chem* 273:11401–11404
78. Jones DP (2008) Radical-free biology of oxidative stress. *Am J Physiol Cell Physiol* 295:C849–C868
79. Halliwell B (2007) Oxidative stress and cancer: have we moved forward? *Biochem J* 401:1–11
80. Thomas DD, Liu X, Kantrow SP, Lancaster JR Jr (2001) The biological lifetime of nitric oxide: implications for the perivascular dynamics of NO and O_2 . *Proc Natl Acad Sci USA* 98:355–360
81. Thomas DD, Ridnour LA, Espey MG, Donzelli S, Ambis S, Hussain SP, Harris CC, DeGraff W, Roberts DD, Mitchell JB, Wink DA (2006) Superoxide fluxes limit nitric oxide-induced signaling. *J Biol Chem* 281:25984–25993
82. Fridovich I (1970) Quantitative aspects of the production of superoxide anion radical by milk xanthine oxidase. *J Biol Chem* 245:4053–4057
83. Jones DP, Eklow L, Thor H, Orrenius S (1981) Metabolism of hydrogen peroxide in isolated hepatocytes: relative contributions of catalase and glutathione peroxidase in decomposition of endogenously generated H_2O_2 . *Arch Biochem Biophys* 210:505–516
84. Go YM, Jones DP (2008) Redox compartmentalization in eukaryotic cells. *Biochim Biophys Acta* 1780(11):1273–1290

85. Barja G (2002) Endogenous oxidative stress: relationship to aging, longevity and caloric restriction. *Ageing Res Rev* 1:397–411
86. Doudican NA, Song B, Shadel GS, Doetsch PW (2005) Oxidative DNA damage causes mitochondrial genomic instability in *saccharomyces cerevisiae*. *Mol Cell Biol* 25:5196–5204
87. Calabrese EJ, Baldwin LA (2003) The hormetic dose-response model is more common than the threshold model in toxicology. *Toxicol Sci* 71:246–250
88. Winterbourn CC, Metodiewa D (1994) The reaction of superoxide with reduced glutathione. *Arch Biochem Biophys* 314:284–290
89. Giles GI, Tasker KM, Jacob C (2001) Hypothesis: the role of reactive sulfur species in oxidative stress. *Free Radic Biol Med* 31:1279–1283
90. Starke DW, Chock PB, Mieyal JJ (2003) Glutathione-thiyl radical scavenging and transferase properties of human glutaredoxin (thioltransferase). Potential role in redox signal transduction. *J Biol Chem* 278:14607–14613
91. Salmeen A, Andersen JN, Myers MP, Meng TC, Hinks JA, Tonks NK, Barford D (2003) Redox regulation of protein tyrosine phosphatase 1B involves a sulphenyl-amide intermediate. *Nature* 423:769–773
92. van Montfort RL, Congreve M, Tisi D, Carr R, Jhoti H (2003) Oxidation state of the active-site cysteine in protein tyrosine phosphatase 1B. *Nature* 423:773–777
93. Biteau B, Labarre J, Toledano MB (2003) Atp-dependent reduction of cysteine-sulphinic acid by *S. cerevisiae* sulphiredoxin. *Nature* 425:980–984
94. Woo HA, Chae HZ, Hwang SC, Yang KS, Kang SW, Kim K, Rhee SG (2003) Reversing the inactivation of peroxiredoxins caused by cysteine sulfinic acid formation. *Science* 300:653–656
95. Woo HA, Kang SW, Kim HK, Yang KS, Chae HZ, Rhee SG (2003) Reversible oxidation of the active site cysteine of peroxiredoxins to cysteine sulfinic acid. Immunoblot detection with antibodies specific for the hyperoxidized cysteine-containing sequence. *J Biol Chem* 278:47361–47364
96. Kim HY, Gladyshev VN (2004) Methionine sulfoxide reduction in mammals: characterization of methionine-r-sulfoxide reductases. *Mol Biol Cell* 15:1055–1064
97. Stadtman ER, Van Remmen H, Richardson A, Wehr NB, Levine RL (2005) Methionine oxidation and aging. *Biochim Biophys Acta* 1703:135–140
98. Gladyshev VN, Hatfield DL (1999) Selenocysteine-containing proteins in mammals. *J Biomed Sci* 6:151–160
99. Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehab O, Guigo R, Gladyshev VN (2003) Characterization of mammalian selenoproteomes. *Science* 300:1439–1443
100. Stadtman ER, Levine RL (2003) Free radical-mediated oxidation of free amino acids and amino acid residues in proteins. *Amino Acids* 25:207–218
101. Carp H, Miller F, Hoidal JR, Janoff A (1982) Potential mechanism of emphysema: alpha 1-proteinase inhibitor recovered from lungs of cigarette smokers contains oxidized methionine and has decreased elastase inhibitory capacity. *Proc Natl Acad Sci USA* 79:2041–2045
102. Hansel A, Kuschel L, Hehl S, Lemke C, Agricola HJ, Hoshi T, Heinemann SH (2002) Mitochondrial targeting of the human peptide methionine sulfoxide reductase (MSRA), an enzyme involved in the repair of oxidized proteins. *FASEB J* 16:911–913
103. Weissbach H, Etienne F, Hoshi T, Heinemann SH, Lowther WT, Matthews B, St John G, Nathan C, Brot N (2002) Peptide methionine sulfoxide reductase: structure, mechanism of action, and biological function. *Arch Biochem Biophys* 397:172–178
104. Miller RA, Buehner G, Chang Y, Harper JM, Sigler R, Smith-Wheelock M (2005) Methionine-deficient diet extends mouse lifespan, slows immune and lens aging, alters glucose, T4, IGF-I and insulin levels, and increases hepatocyte MIF levels and stress resistance. *Aging Cell* 4:119–125
105. Moskovitz J, Bar-Noy S, Williams WM, Requena J, Berlett BS, Stadtman ER (2001) Methionine sulfoxide reductase (MSRA) is a regulator of antioxidant defense and lifespan in mammals. *Proc Natl Acad Sci USA* 98:12920–12925
106. Richie JP Jr, Leutzinger Y, Parthasarathy S, Malloy V, Orentreich N, Zimmerman JA (1994) Methionine restriction increases blood glutathione and longevity in F344 rats. *FASEB J* 8:1302–1307
107. Sanz A, Caro P, Ayala V, Portero-Otin M, Pamplona R, Barja G (2006) Methionine restriction decreases mitochondrial oxygen radical generation and leak as well as oxidative damage to mitochondrial DNA and proteins. *FASEB J* 20:1064–1073
108. Jones DP, Go YM (2010) Mapping the cysteine proteome: analysis of redox-sensing thiols. *Curr Opin Chem Biol* 15:1–10
109. Hansen JM, Go YM, Jones DP (2006) Nuclear and mitochondrial compartmentation of oxidative stress and redox signaling. *Annu Rev Pharmacol Toxicol* 46:215–234
110. Gallogly MM, Starke DW, Mieyal JJ (2009) Mechanistic and kinetic details of catalysis of thiol-disulfide exchange by glutaredoxins and potential mechanisms of regulation. *Antioxid Redox Signal* 11:1059–1081
111. Kemp M, Go YM, Jones DP (2008) Nonequilibrium thermodynamics of thiol/disulfide redox systems: a perspective on redox systems biology. *Free Radic Biol Med* 44:921–937
112. Cenci S, Sitia R (2007) Managing and exploiting stress in the antibody factory. *FEBS Lett* 581:3652–3657
113. Todd DJ, Lee AH, Glimcher LH (2008) The endoplasmic reticulum stress response in immunity and autoimmunity. *Nat Rev Immunol* 8:663–674
114. Go YM, Jones DP (2010) Redox control systems in the nucleus: mechanisms and functions. *Antioxid Redox Signal* 13:489–509
115. Goldberger AL, Amaral LA, Hausdorff JM, Ivanov P, Peng CK, Stanley HE (2002) Fractal dynamics in physiology: alterations with disease and aging. *Proc Natl Acad Sci USA* 99(Suppl 1):2466–2472

116. Ivanov PC, Amaral LA, Goldberger AL, Havlin S, Rosenblum MG, Struzik ZR, Stanley HE (1999) Multifractality in human heartbeat dynamics. *Nature* 399:461–465
117. Greenberg M (2009) Radical and radical ion reactivity in nucleic acid chemistry. Wiley, New York, 466 pp
118. Sarkar D, Datta R, Hannigan R (2007) Concepts and applications in environmental geochemistry. Elsevier, Amsterdam, 761 pp
119. Keilin D (1966) The history of cell respiration and cytochrome. In: Keilin D (ed) The history of cell respiration and cytochrome. Cambridge University Press, Cambridge
120. Finkel T (2003) Oxidant signals and oxidative stress. *Curr Opin Cell Biol* 15:247–254
121. Wu JJ, Quijano C, Chen E, Liu H, Cao L, Fergusson MM, Rovira II, Gutkind S, Daniels MP, Komatsu M, Finkel T (2009) Mitochondrial dysfunction and oxidative stress mediate the physiological impairment induced by the disruption of autophagy. *Aging* 1:425–437
122. Swanson RA, Siesjo B (2002) Oxidative stress in cerebrovascular disease. In: Swanson RA, Siesjo B (eds) Oxidative stress in cerebrovascular disease. Cambridge University Press, Cambridge